

VisTrails: Provenance and Data

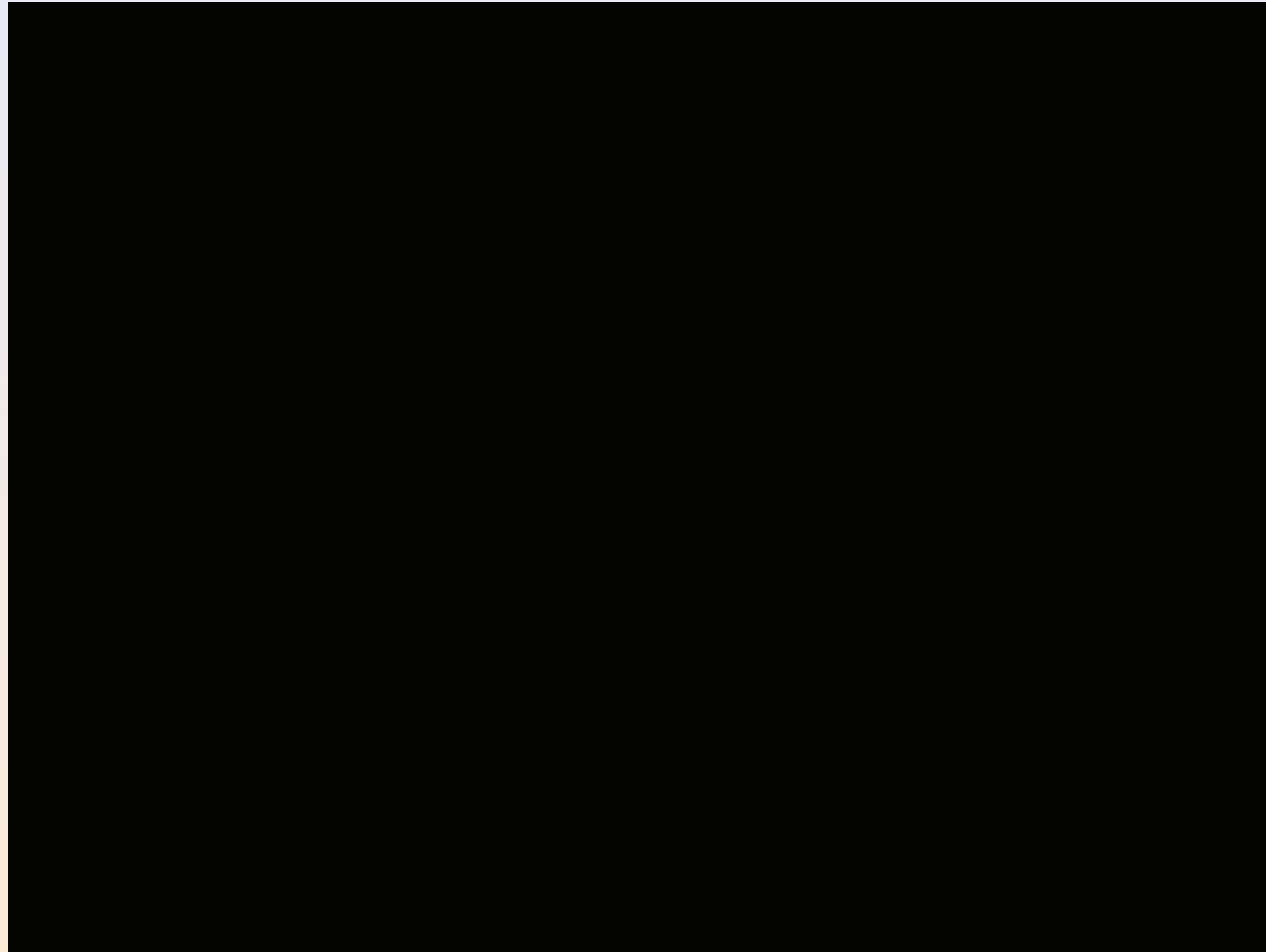
Cláudio T. Silva

Scientific Computing and Imaging Institute
School of Computing
University of Utah

Introduction

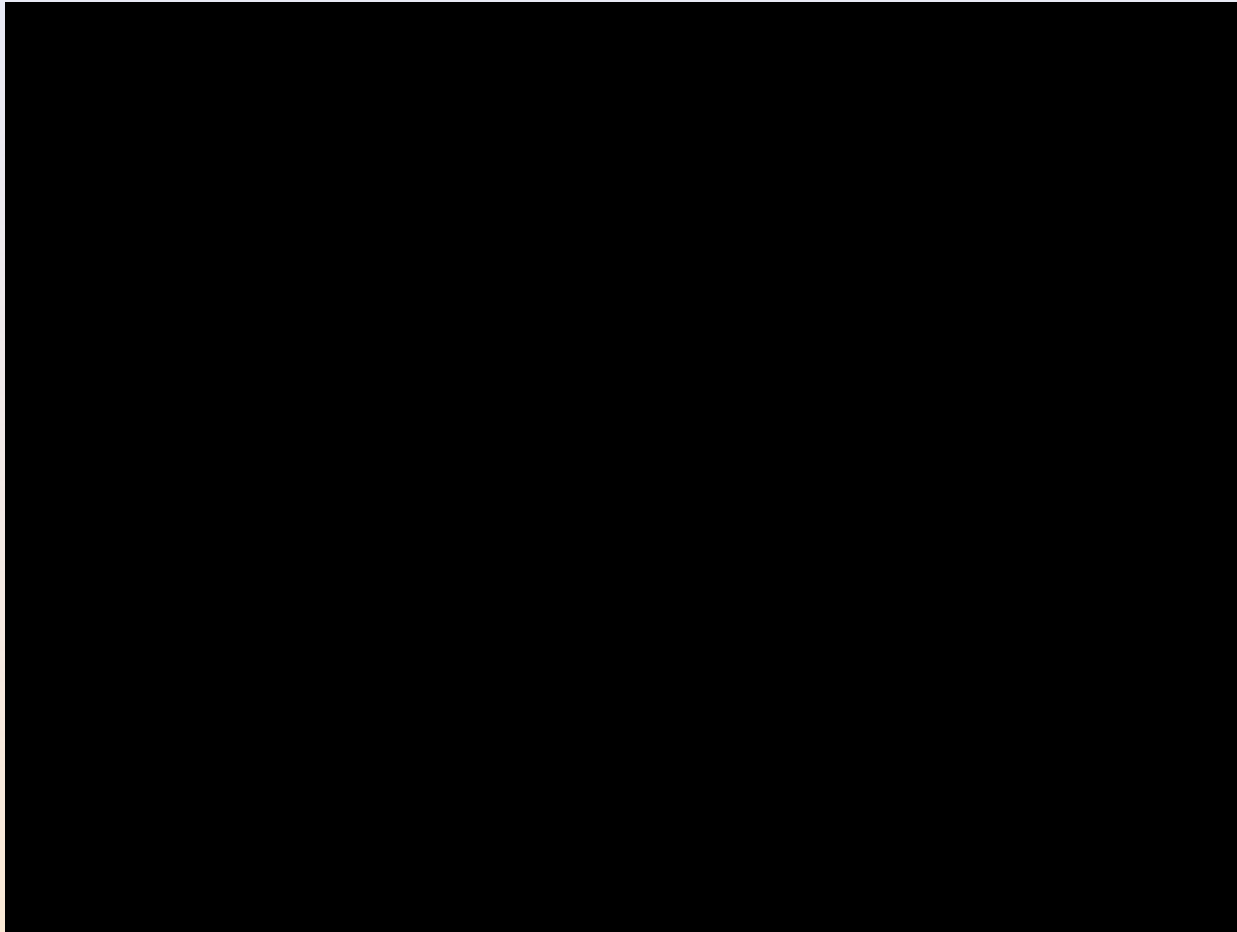
- ◆ Presentation and Exploratory VIS
- ◆ Influential VIS tools
- ◆ Extending VIS tools to facilitate discovery
- ◆ Ongoing (computer science) research
- ◆ Conclusions

Presentation VIS



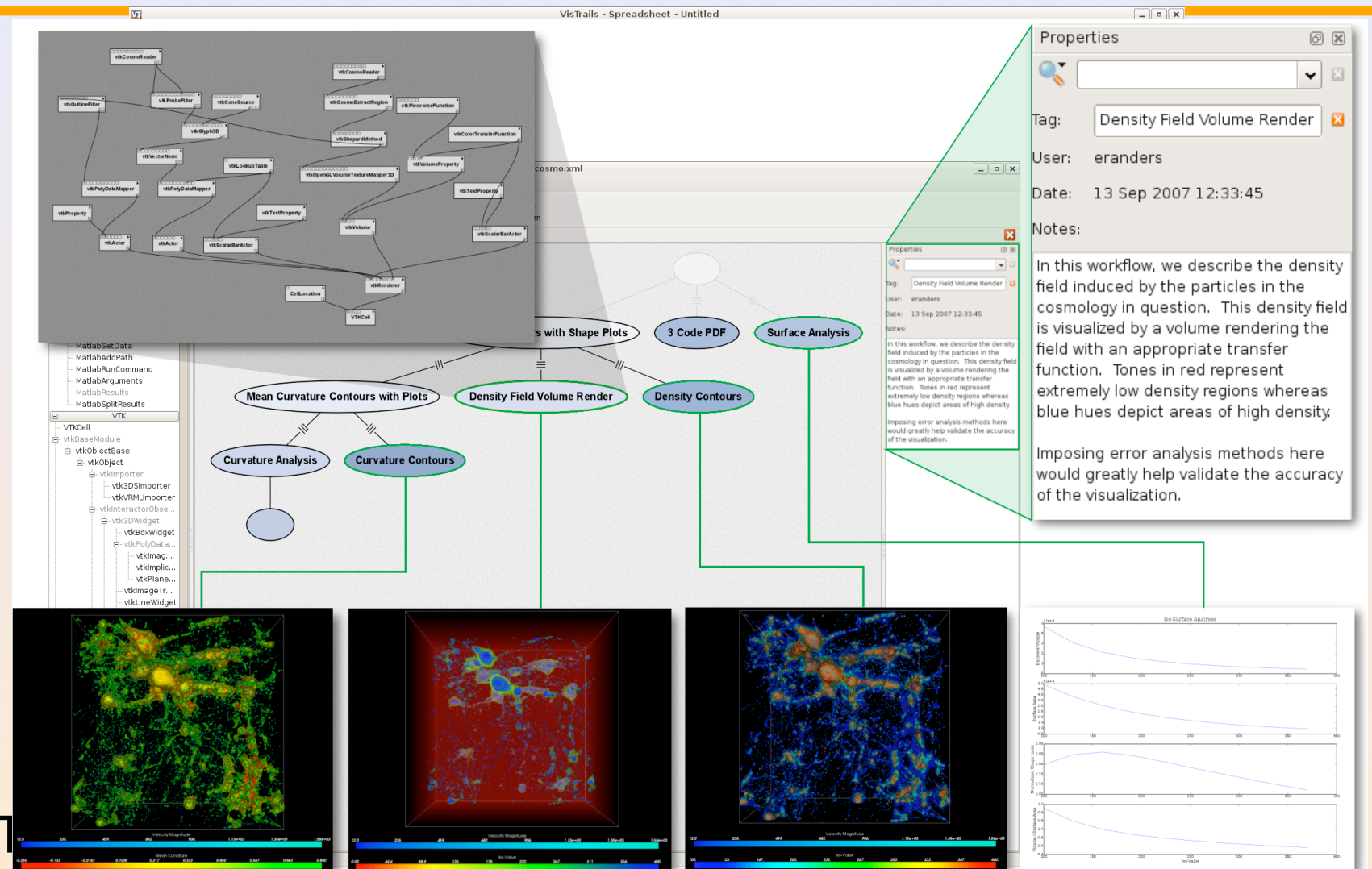
“Study of a Numerically Modeled Severe Storm”, NCSA, UIUC

Presentation VIS



“Fusion Simulation Visualization” Kruger, Sanderson, et al

Exploratory VIS



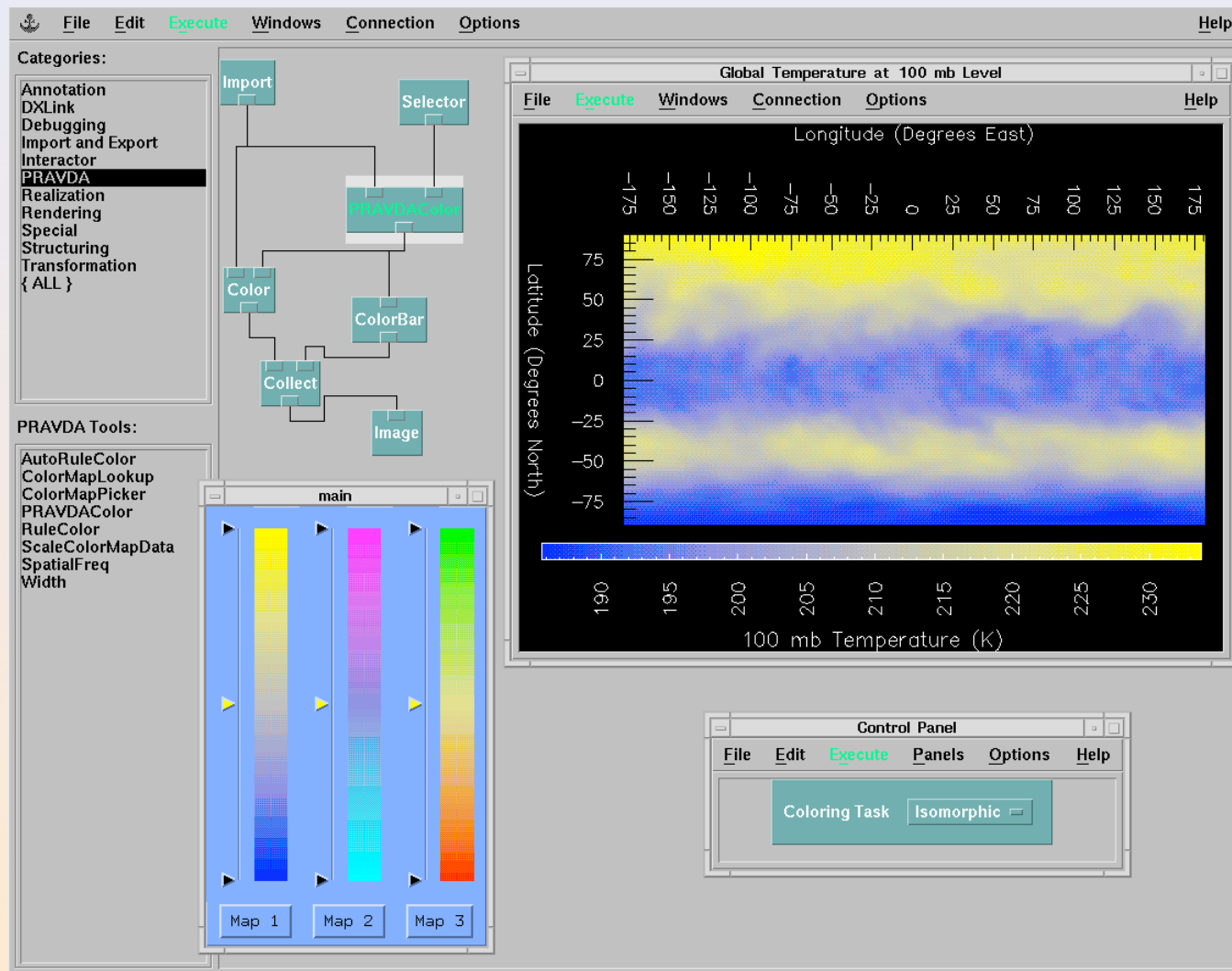
Anderson, Heitmann, Habib, et al

Most Popular Computer Tool for “Exploratory Discovery”?

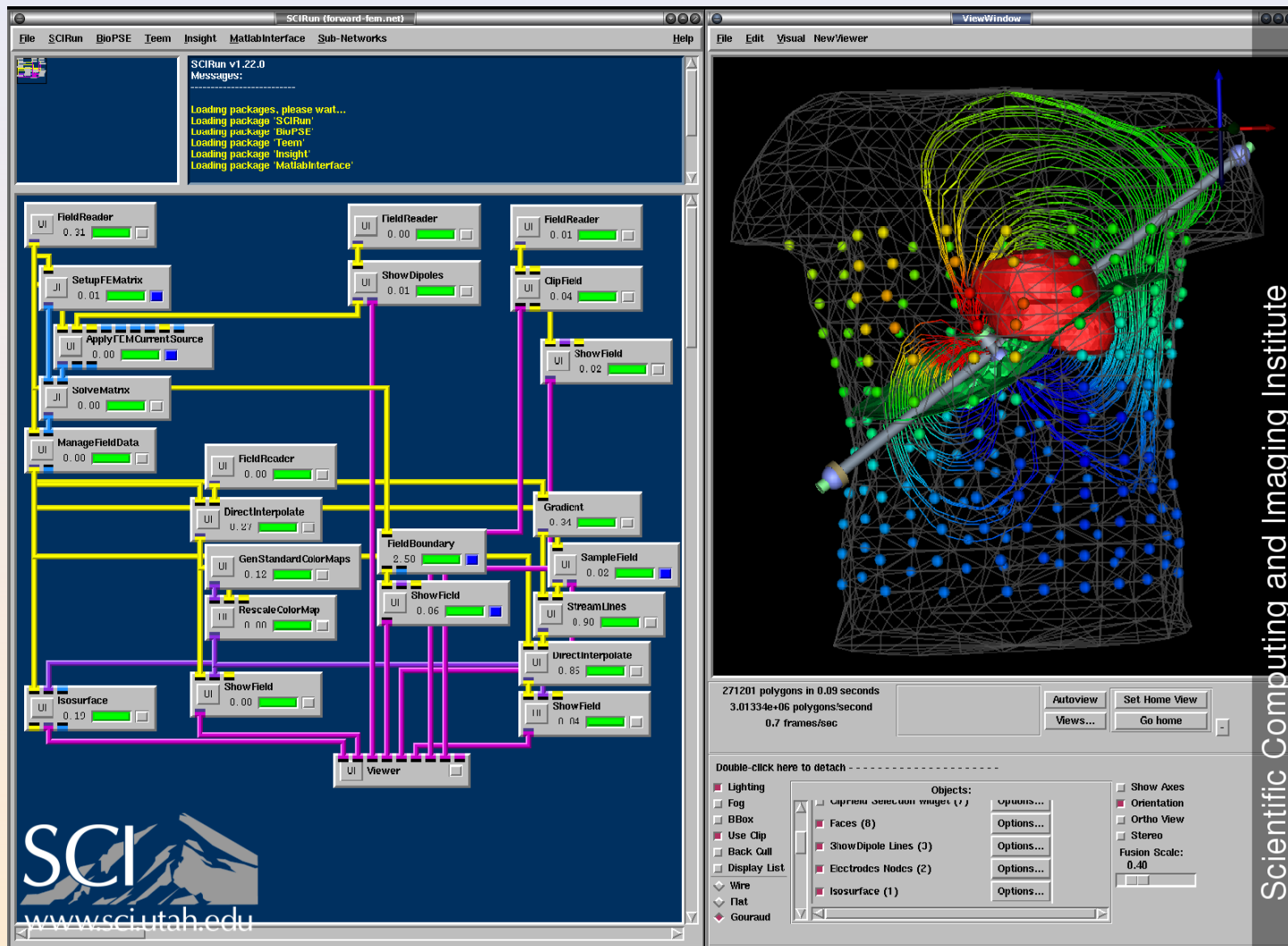
- ◆ Microsoft PowerPoint!

Influential VIS tools

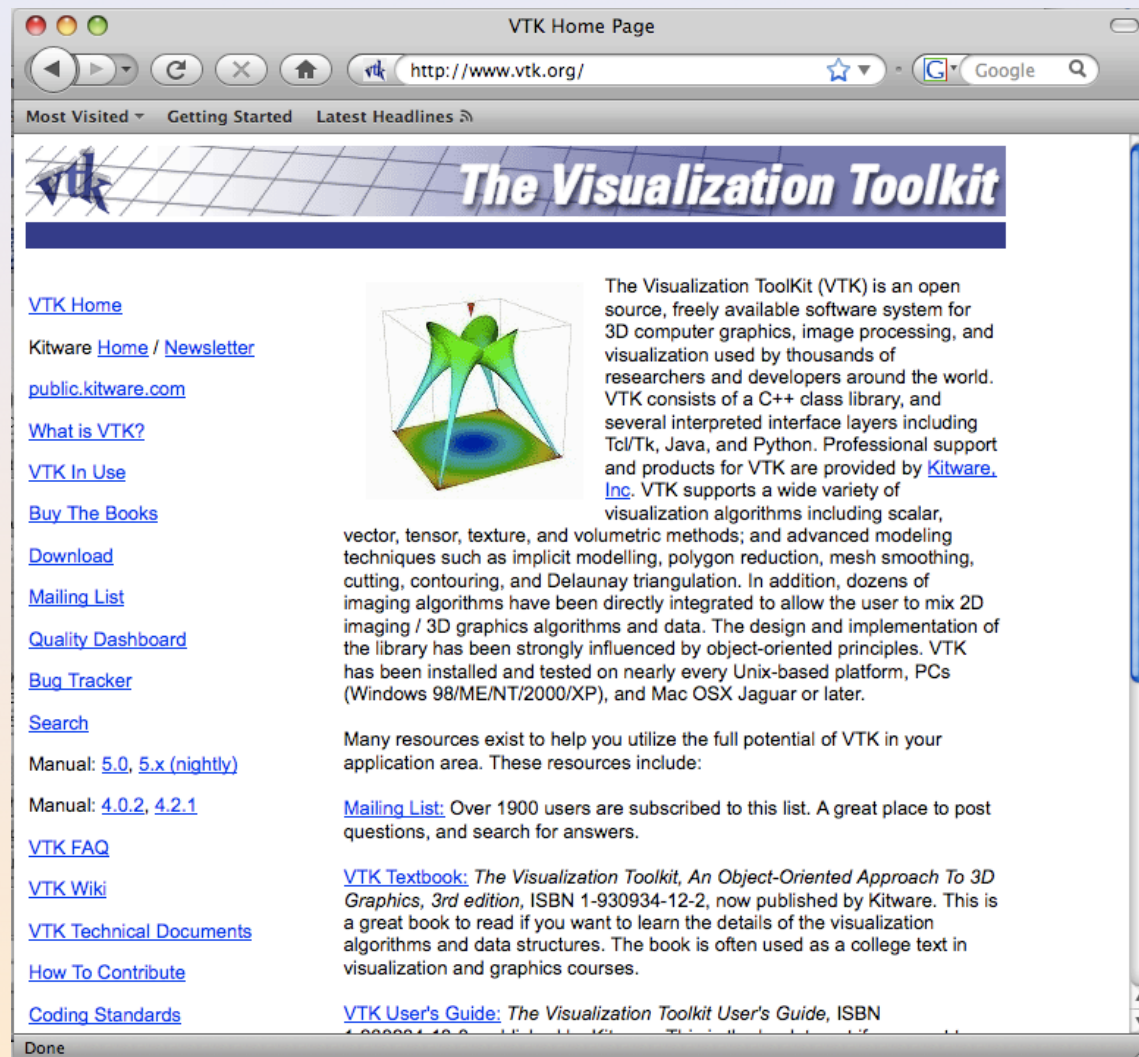
IBM OpenDX



SCIRun



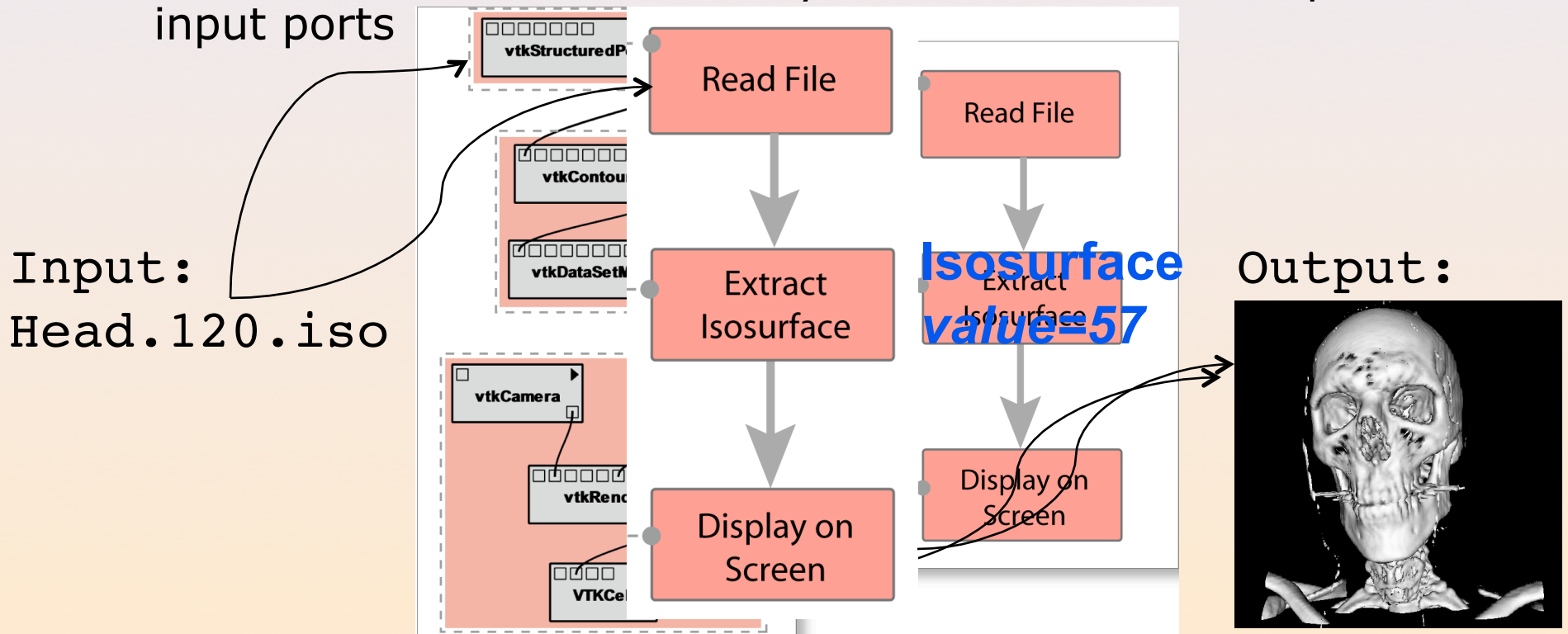
VTK



Digression: Workflows

Scientific Workflows and Dataflows

- ◆ Dataflows are directed graphs describing a computational task
 - Vertices = modules = processing steps + **parameters**
 - Edges = connections between output and input ports
 - Execution order determined by flow of data from output to input ports



Scientific Workflows and Dataflows

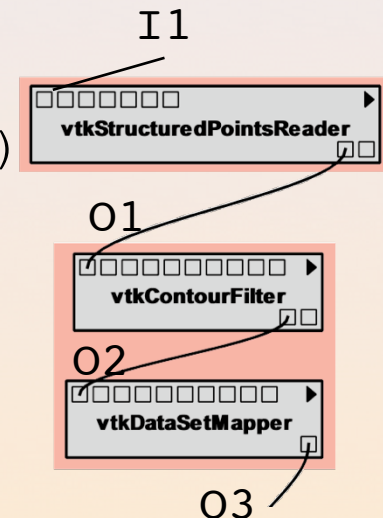
- ◆ A directed graph describing a computational task
 - Vertices = modules = processing steps + parameters
 - Edges = connections between output and input ports
 - Execution order determined by flow of data from output to input ports
- ◆ No state or side effects: Outputs are a *function* of the inputs

```
O3 = vtkDataSetMapper(input=O2)
```

```
O2 = vtkContourFilter(value=57,input=O1)
```

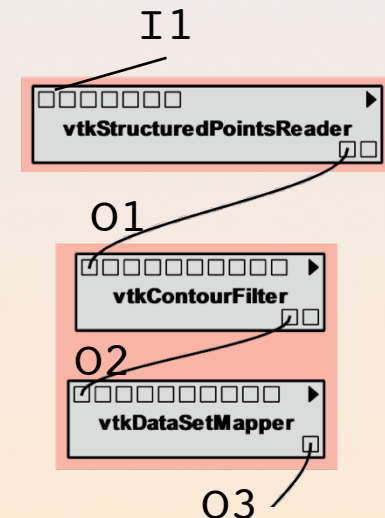
```
O1 = vtkStructuredReader(input=I1)
```

[Lee and Parks, IEEE 1995]



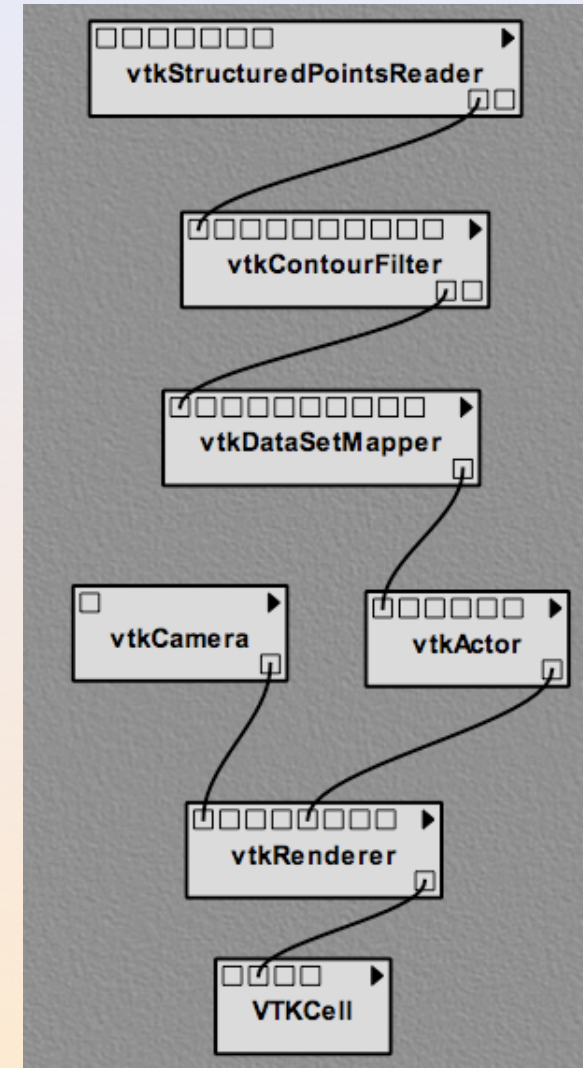
Scientific Workflows and Dataflows

- ◆ A directed graph describing a computational task
 - Vertices = modules = processing steps + parameters
 - Edges = connections between output and input ports
 - Execution order determined by flow of data from output to input ports
- ◆ No state or side effects: Outputs are a *function* of the inputs
- ◆ Simple programming model
 - Good match for visual programming interfaces
 - Widely used: adopted by most scientific workflow and visualization systems
 - Easy to optimize and parallelize



Workflows and Computer Programs

```
1 import vtk
2
3 data = vtk.vtkStructuredPointsReader()
4 data.SetFileName("../examples/data/head.120.vtk")
5
6 contour = vtk.vtkContourFilter()
7 contour.SetInput(0, data.GetOutput())
8 contour.SetValue(0, 67)
9
10 mapper = vtk.vtkPolyDataMapper()
11 mapper.SetInput(contour.GetOutput())
12 mapper.ScalarVisibilityOff()
13
14 actor = vtk.vtkActor()
15 actor.SetMapper(mapper)
16
17 cam = vtk.vtkCamera()
18 cam.SetViewUp(0, 0, -1)
19 cam.SetPosition(745, -453, 369)
20 cam.SetFocalPoint(135, 135, 150)
21 cam.ComputeViewPlaneNormal()
22
23 ren = vtk.vtkRenderer()
24 ren.AddActor(actor)
25 ren.SetActiveCamera(cam)
26 ren.ResetCamera()
27
28 renwin = vtk.vtkRenderWindow()
29 renwin.AddRenderer(ren)
30
31 style = vtk.vtkInteractorStyleTrackballCamera()
32 iren = vtk.vtkRenderWindowInteractor()
33 iren.SetRenderWindow(renwin)
34 iren.SetInteractorStyle(style)
35 iren.Initialize()
36 iren.Start()
```



Workflows and Computer Programs

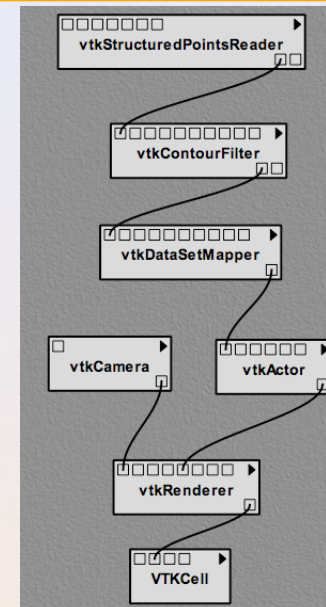
```
1 import vtk
2
3 data = vtk.vtkStructuredPointsReader()
4 data.SetFileName("../examples/data/head.128.vtk")
5
6 contour = vtk.vtkContourFilter()
7 contour.SetInput(0, data.GetOutput())
8 contour.SetValue(0, 67)
9
10 mapper = vtk.vtkPolyDataMapper()
11 mapper.SetInput(contour.GetOutput())
12 mapper.ScalarVisibilityOff()
13
14 actor = vtk.vtkActor()
15 actor.SetMapper(mapper)
16
17 cam = vtk.vtkCamera()
18 cam.SetViewUp(0, 0, -1)
19 cam.SetPosition(745, -453, 369)
20 cam.SetFocalPoint(135, 135, 150)
21 cam.ComputeViewPlaneNormal()
22
23 ren = vtk.vtkRenderer()
24 ren.AddActor(actor)
25 ren.SetActiveCamera(cam)
26 ren.ResetCamera()
27
28 renwin = vtk.vtkRenderWindow()
29 renwin.AddRenderer(ren)
30
31 style = vtk.vtkInteractorStyleTrackballCamera()
32 iren = vtk.vtkRenderWindowInteractor()
33 iren.SetRenderWindow(renwin)
34 iren.SetInteractorStyle(style)
35 iren.Initialize()
36 iren.Start()
```

Program

Workflow

Document

Database



2. [The Advanced Html Companion](#)
by Keith Schengili-Roberts, Kim Silk-Copeland. Paperback (August 1998)
Our Price: \$35.96
You Save: \$8.99 (20%)
Usually ships in 24 hours
Average Customer Review: ★★★★★
3. [Applied XML Solutions \(Sams Professional Publishing\)](#)
by Benoit Marchal. Paperback (August 29, 2000)
Our Price: \$35.99
You Save: \$9.00 (20%)
Usually ships in 24 hours
Average Customer Review: ★★★★★
4. [Applied XML: A Toolkit for Programmers](#)
by Alex Ceponkus, Faraz Hoodbhoy. Paperback (July 1, 1999)
Our Price: \$39.99
You Save: \$10.00 (20%)
Usually ships in 24 hours
Average Customer Review: ★★★★★

<Book>

<Title>The Advanced Html Companion</Title>

<Author> Keith Schengili-Roberts </Author>

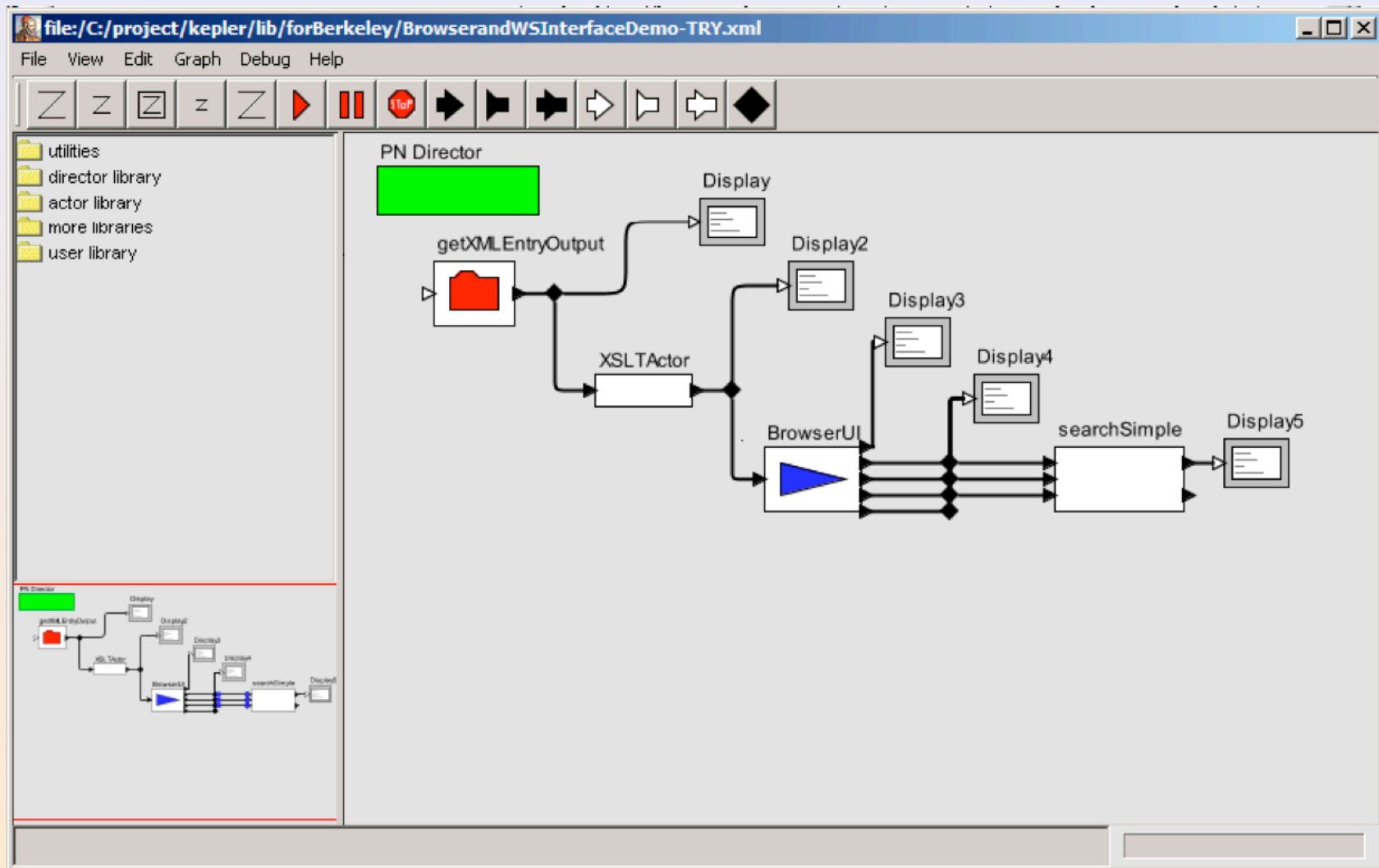
<Author> Kim Silk-Copeland</Author>

<Price> 35.96</Price>...

</Book>

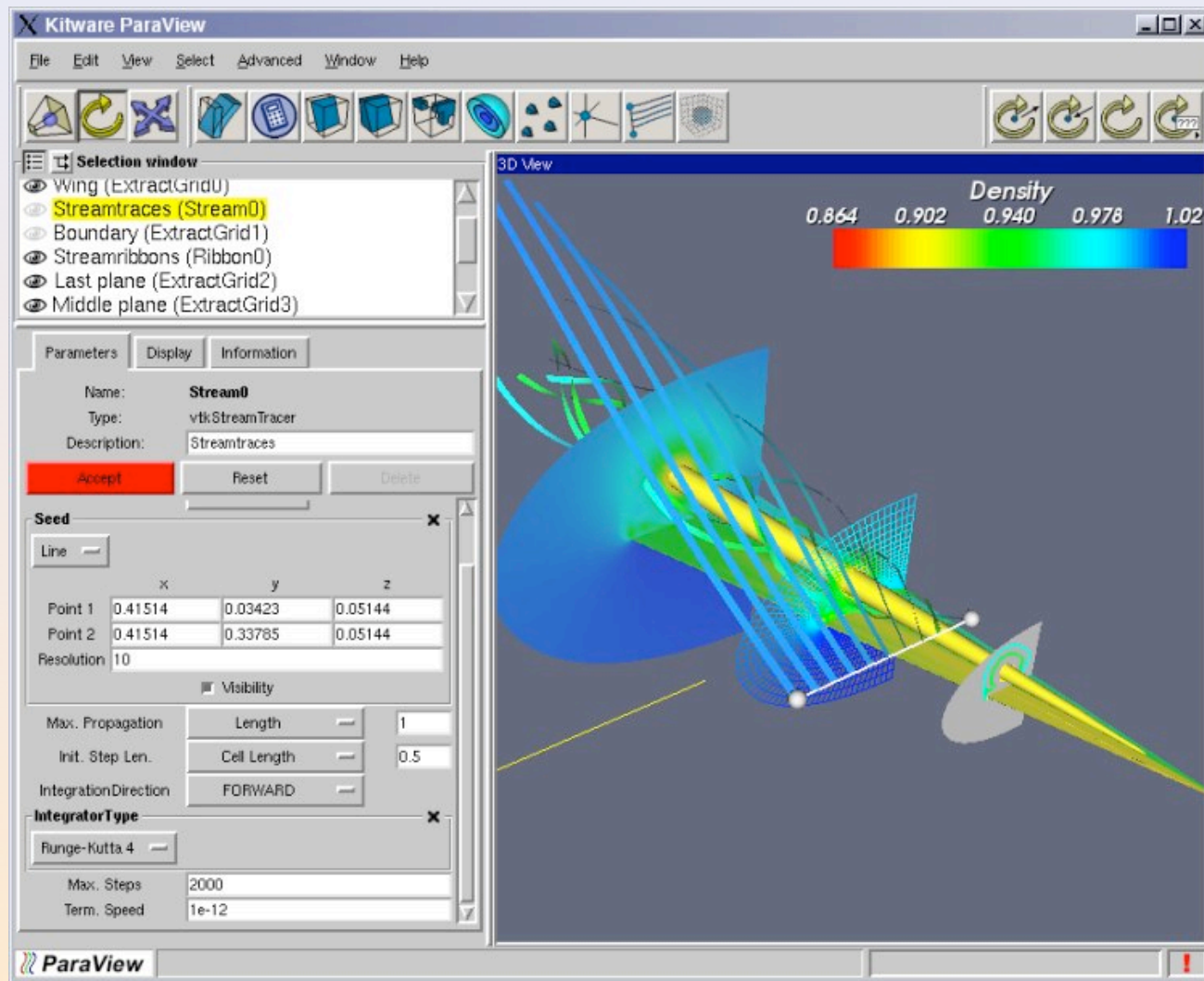
A program is to a workflow what an unstructured document is to a (structured) database.

Kepler

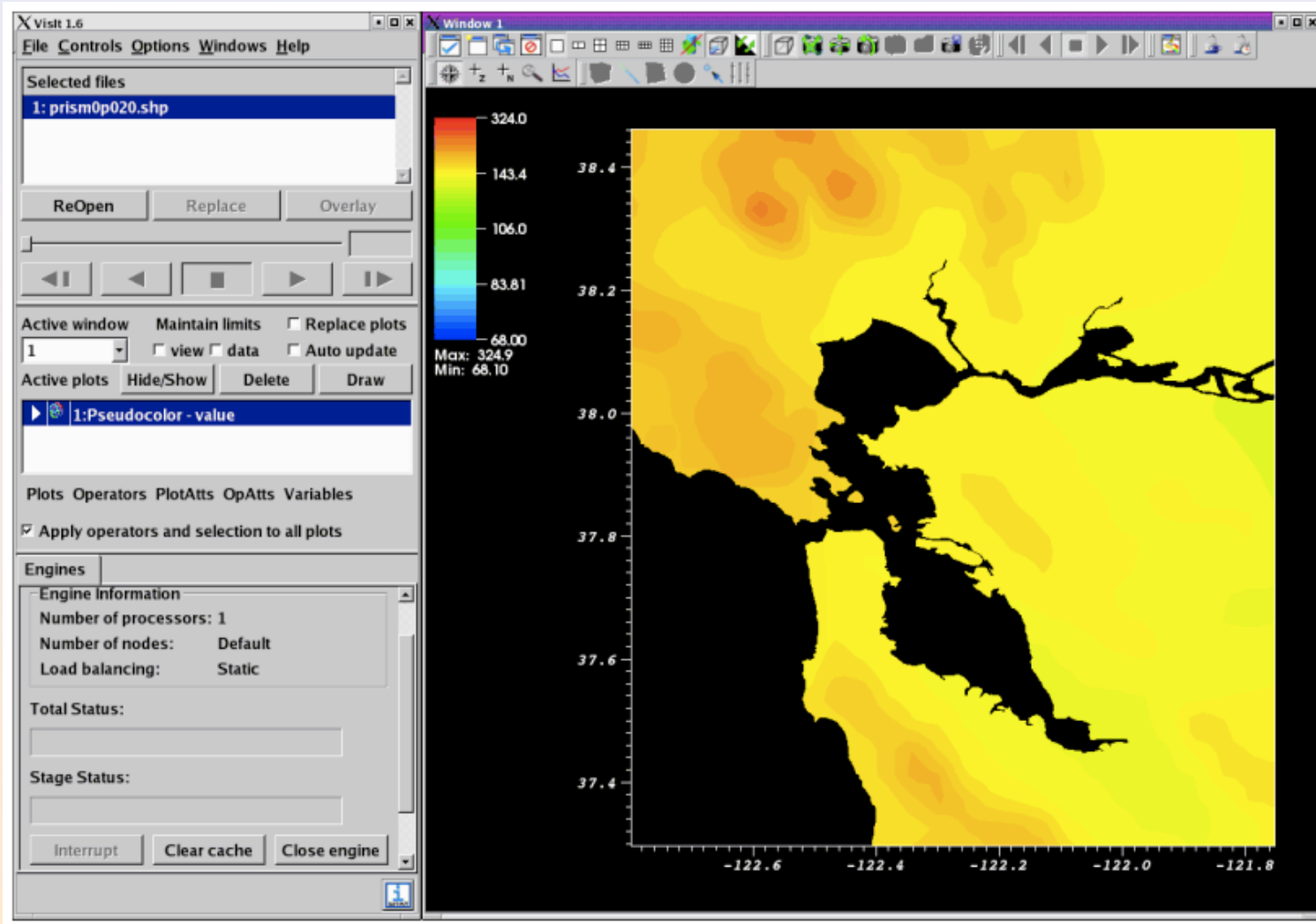


Back to VIS

ParaView



VisIt



VisTrails Project

Exploration and Workflows

- ◆ Workflows have been traditionally used to automate repetitive tasks
- ◆ In exploratory tasks, *change is the norm!*
 - Data analysis and exploration are iterative processes

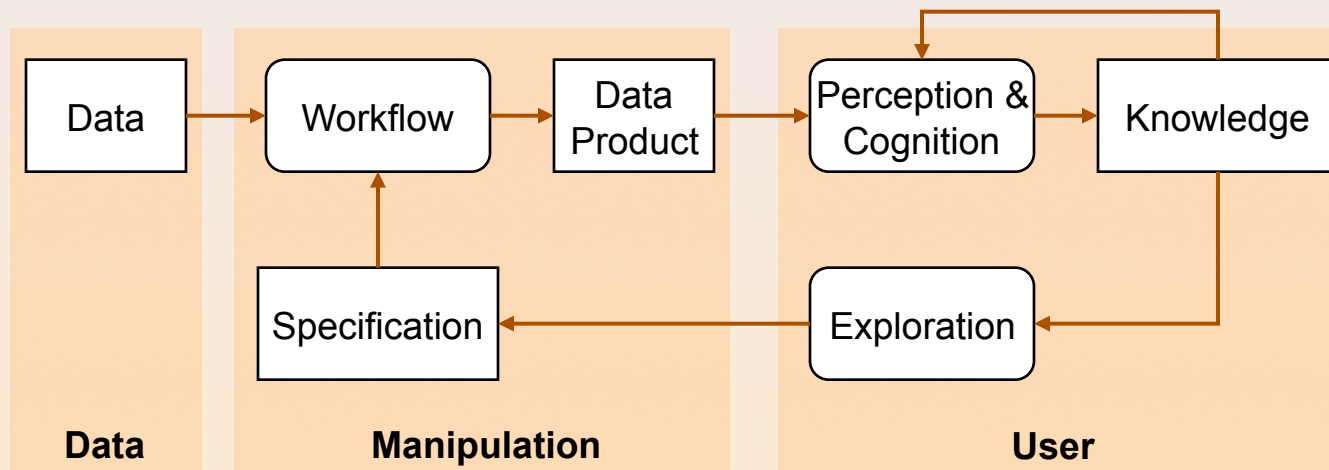


Figure modified from J. van Wijk, IEEE Vis 2005

Exploration and Creativity Support

- ◆ Reflective reasoning is key in the exploratory processes
- ◆ *"Reflective reasoning requires the ability to store temporary results, to make inferences from stored knowledge, and to follow chains of reasoning backward and forward, sometimes backtracking when a promising line of thought proves to be unfruitful. ...the process is slow and laborious"*

Donald A. Norman

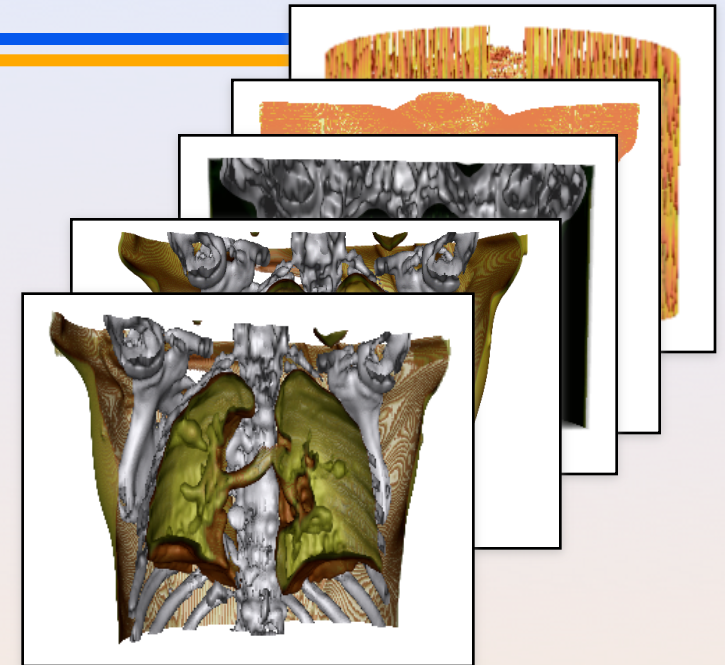
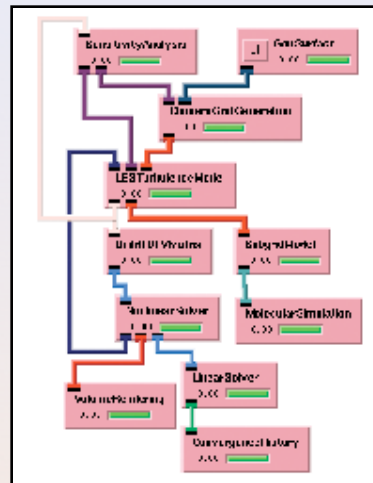
- ◆ Need external aids—tools to facilitate this process
 - Creativity support tools [Shneiderman, CACM 2002]
- ◆ Need aid from people—collaboration

Data Exploration and Workflows

raw data:CT scan



workflow



Files (workflow specifications)

anon4877_voxel_scale_1_zspace_20060331.srn

anon4877_textureshading_20060331.srn

anon4877_textureshading_plane0_20060331.srn

anon4877_goodxferfunction_20060331.srn

anon4877_lesion_20060331.srn

Notes

Initial
visualization

Added texture

an Added plane to

Found good

Identified
lesion tissue

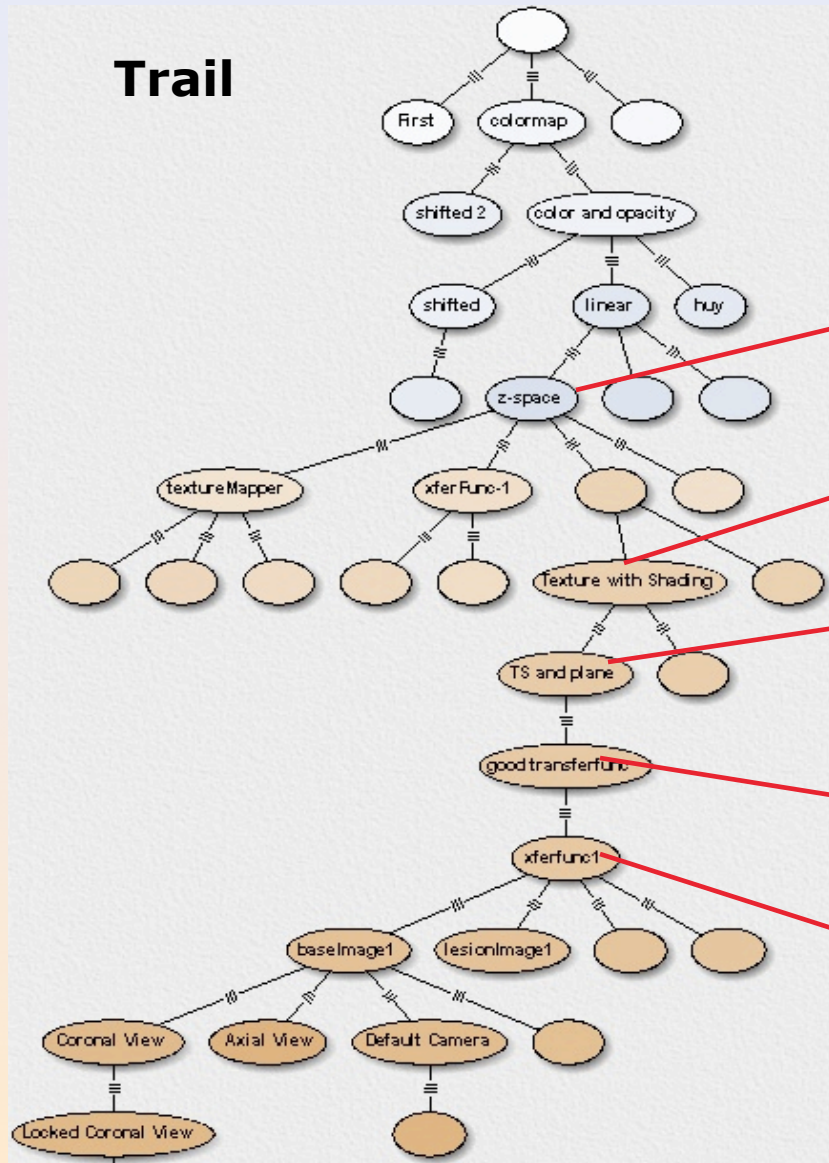
VisTrails: Managing Exploration

- ◆ Comprehensive *provenance infrastructure* for computational tasks
 - Data + **workflow** provenance
 - *Treat workflow as a 1st-class data product*
- ◆ Support for *exploratory* tasks such as visualization and data mining
 - Task specification iteratively refined as users generate and test hypotheses
- ◆ VisTrails **manages the data, metadata and the exploration process**, scientists can focus on *science!*
- ◆ Not a replacement for visualization or scientific workflow systems: infrastructure that can be combined with and enhance these systems
- ◆ **Focus on usability**—build tools for scientists

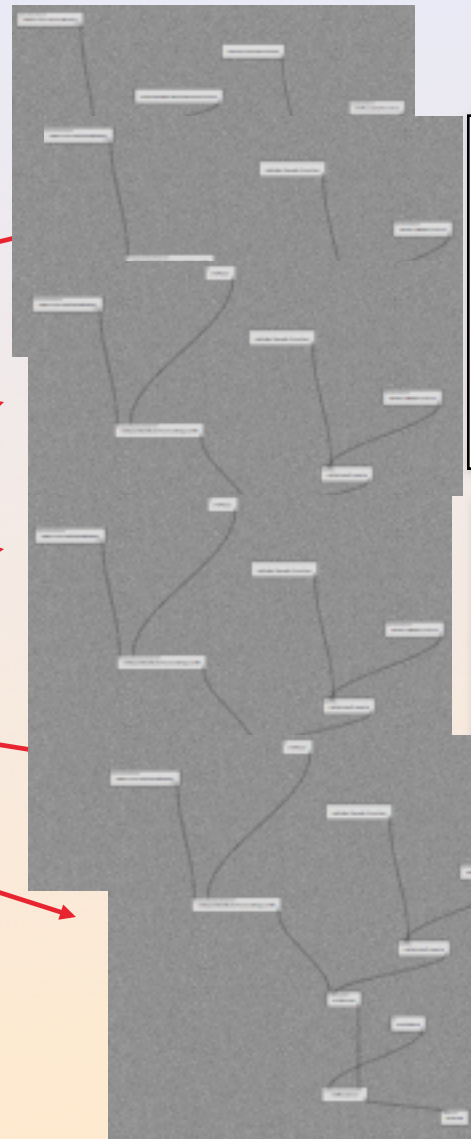
<http://www.vistrails.org>

Keeping Exploration Trails

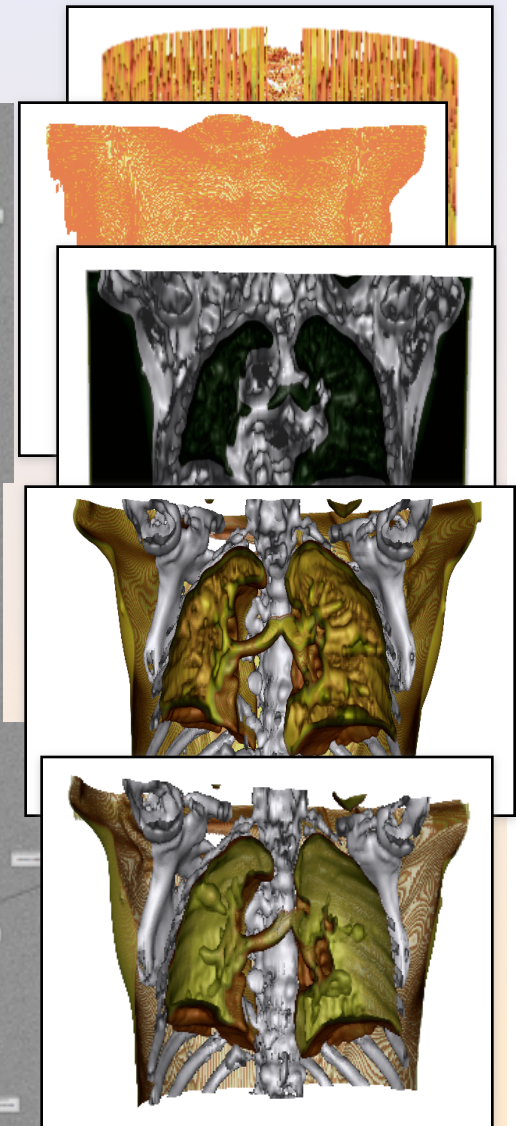
Trail



Workflows



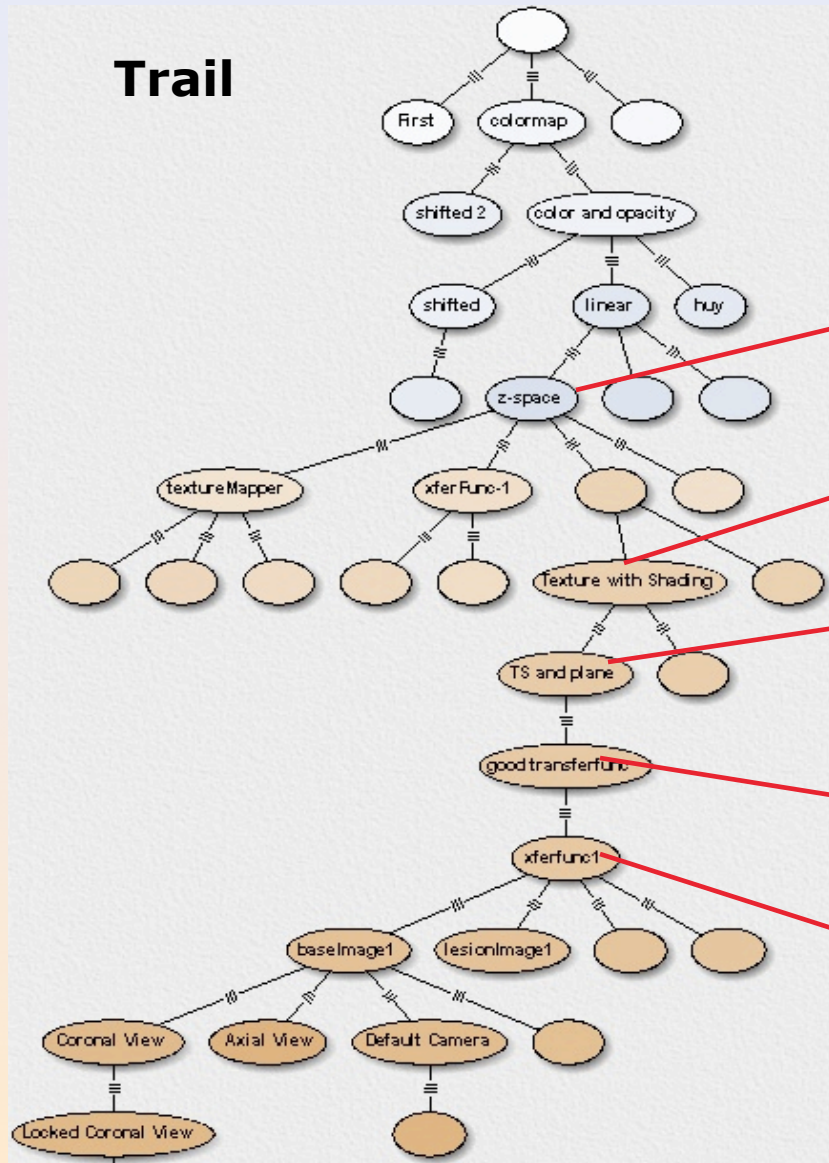
Data Products



Software for Exploratory Visualization

Keeping Exploration Trails

Trail



Notes

User

Initial
visualization
with z
-scaling
corrected

juliana

Added texture
and shading

eranders

Added plane to visualize
internal structure

eranders

Found good
transfer
function

eranders

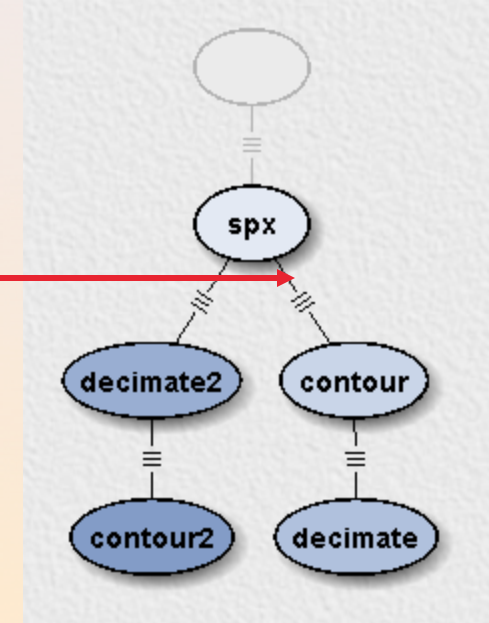
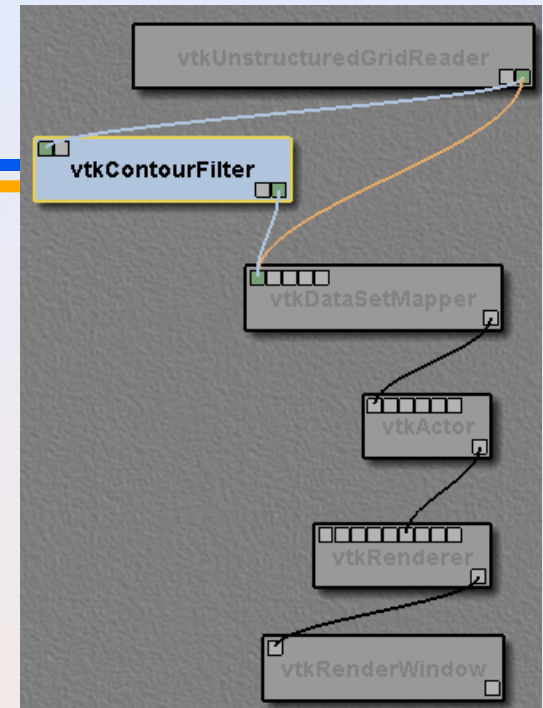
Identified
lesion tissue

stevec

Change-Based Provenance

- ◆ Records actions
- ◆ Provenance = changes to computational tasks
 - Add a module, add a connection, change a parameter value
- ◆ Extensible *change* algebra

addModule
deleteConnection
addConnection
addConnection
setParameter



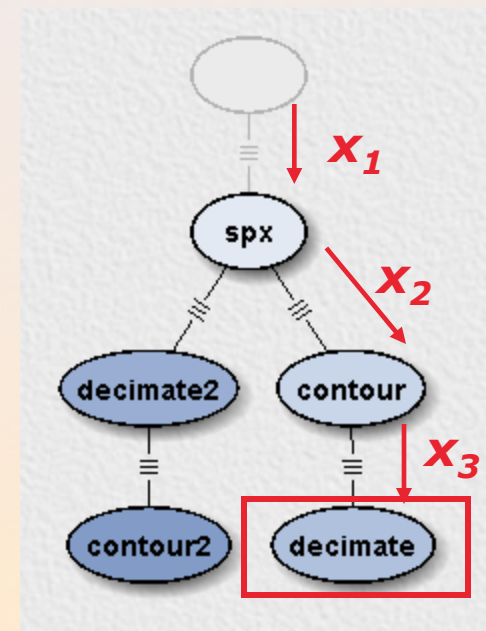
Change-Based Provenance

- ◆ Records actions
- ◆ Provenance = changes to computational tasks
 - Add a module, add a connection, change a parameter value
- ◆ Extensible *change* algebra
- ◆ A *vistrail* node v_t corresponds to the workflow that is constructed by the sequence of actions from the root to v_t

$$V_t = X_n \circ X_{n-1} \circ \dots \circ X_1 \circ \emptyset$$

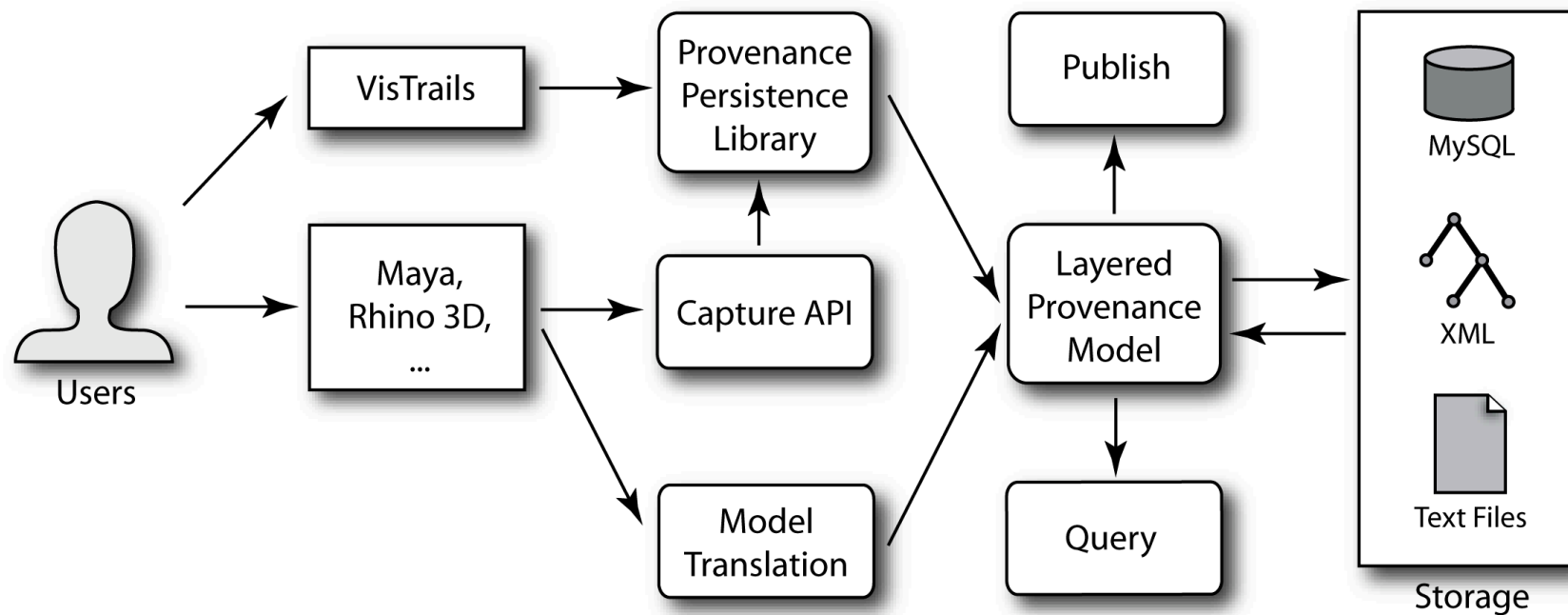
[Freire et al, IPAW 2006]

vistrail

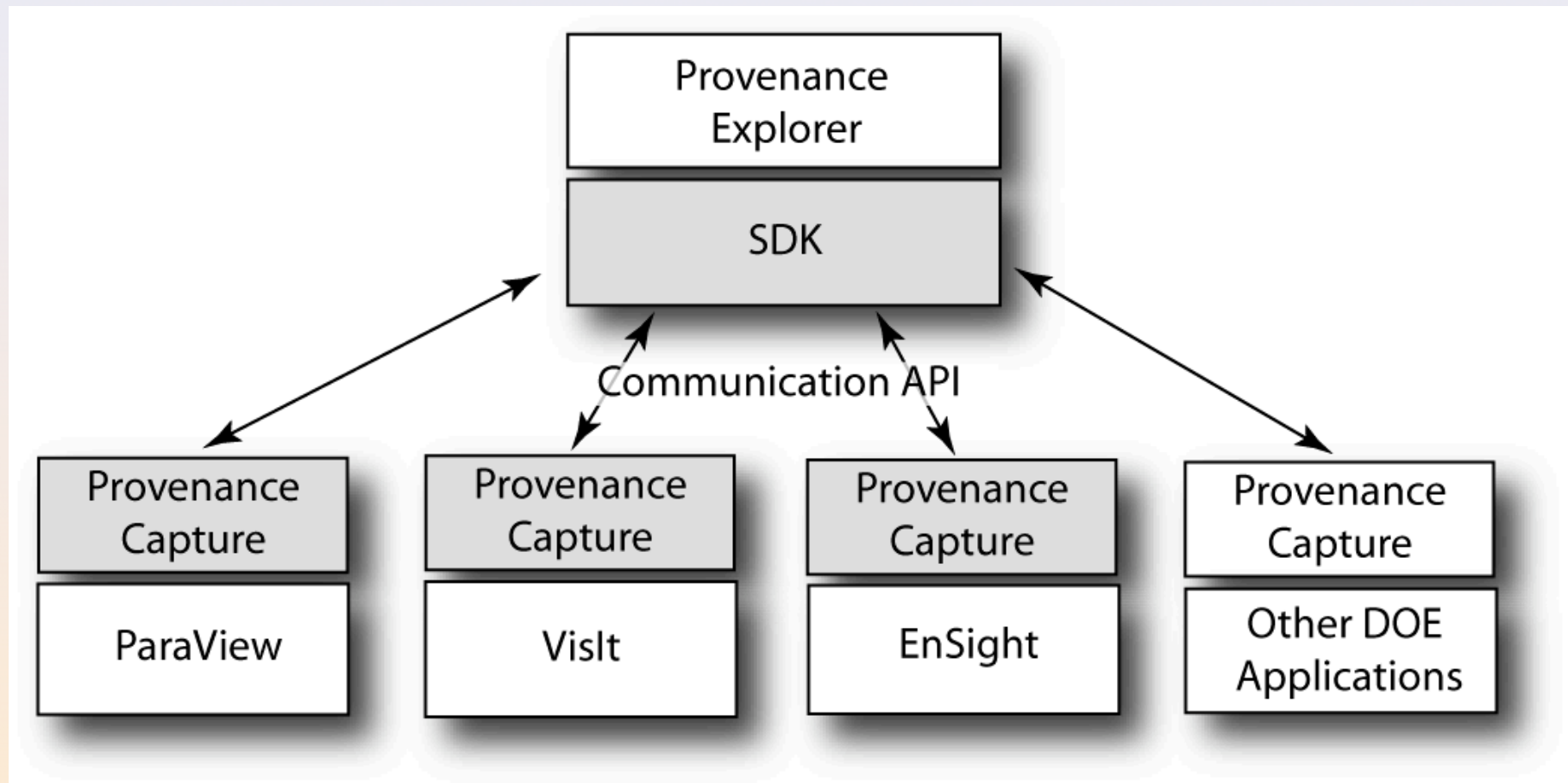




Provenance API

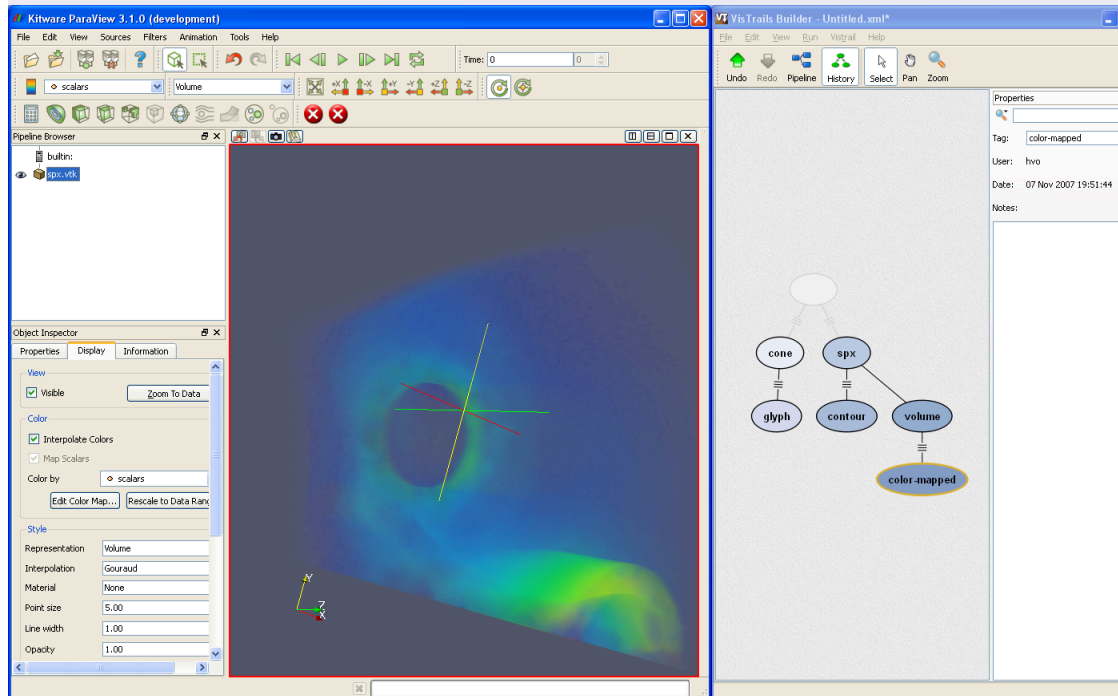


Provenance “Plug-ins”



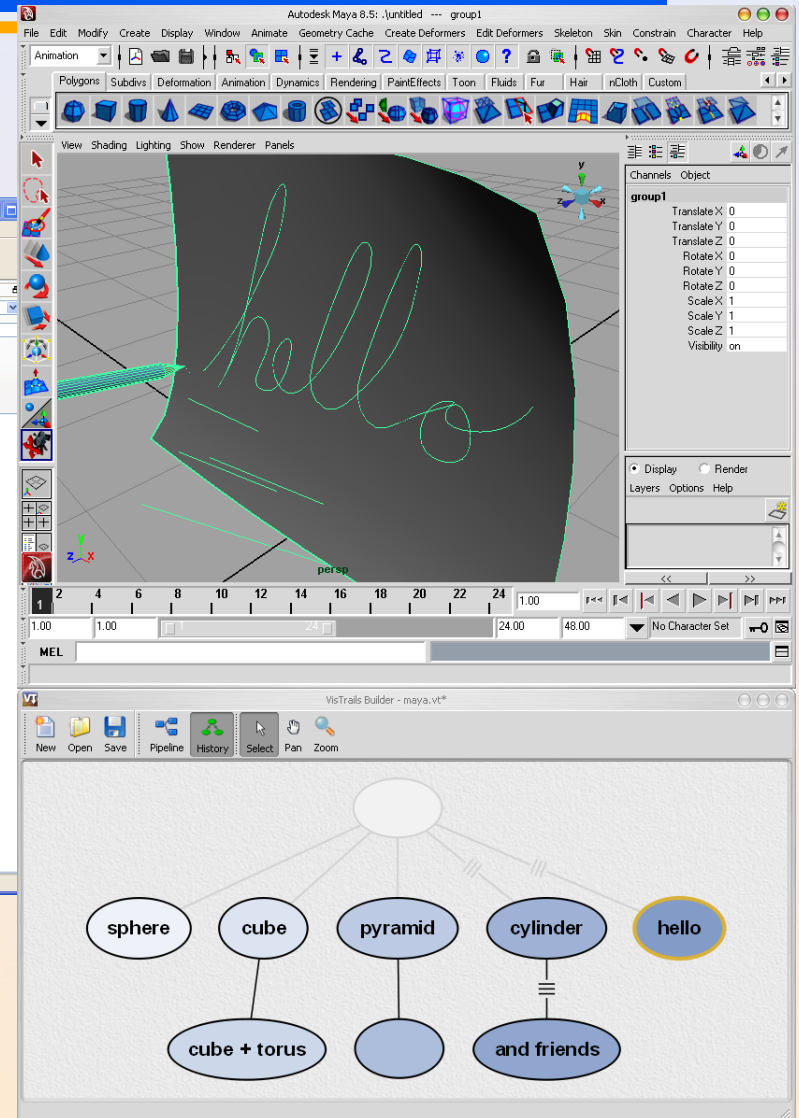
Provenance Enabling 3rd-Party Tools

VisTrails add-on for ParaView



[Callahan et al., IPAW 2008]

Software for Exploratory Visualization



VisTrails add-on for Maya
Silva & Freire

33

ParaView (video)

Provenance Explorer Plug-in
for
Kitware's ParaView 3.0

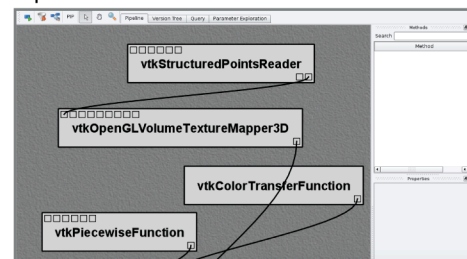
VisTrails Inc.

Sample of Ongoing Work

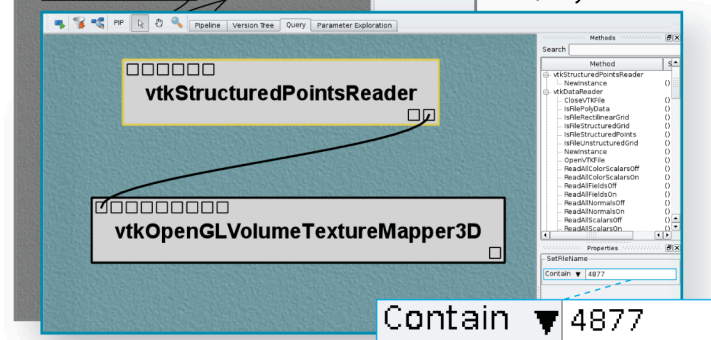
Querying Workflows by Example

- ◆ Workflows are graphs: hard to specify queries using text!
- ◆ Querying workflows by example [Scheidegger et al., TVCG 2007; Beeri et al., VLDB 2006; Beeri et al. VLDB 2007]
 - WYSIWYQ -- What You See Is What You Query
 - Interface to create workflow is same as to query

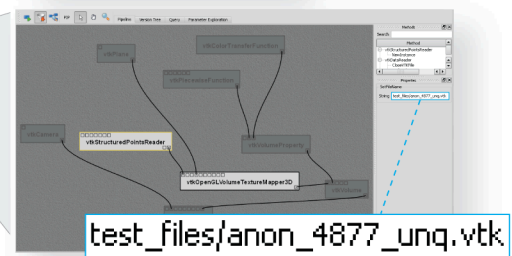
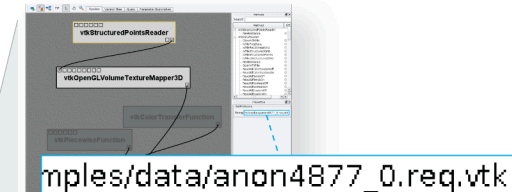
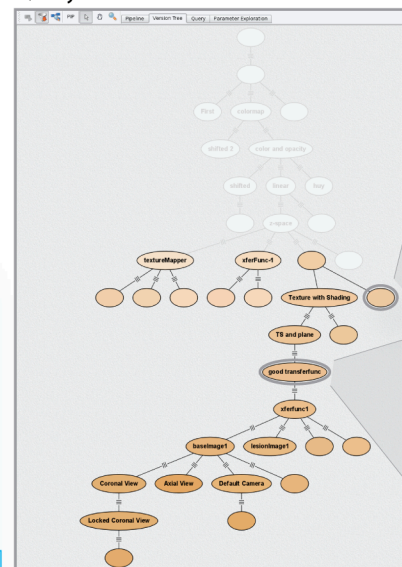
Pipeline Interface



Query Interface



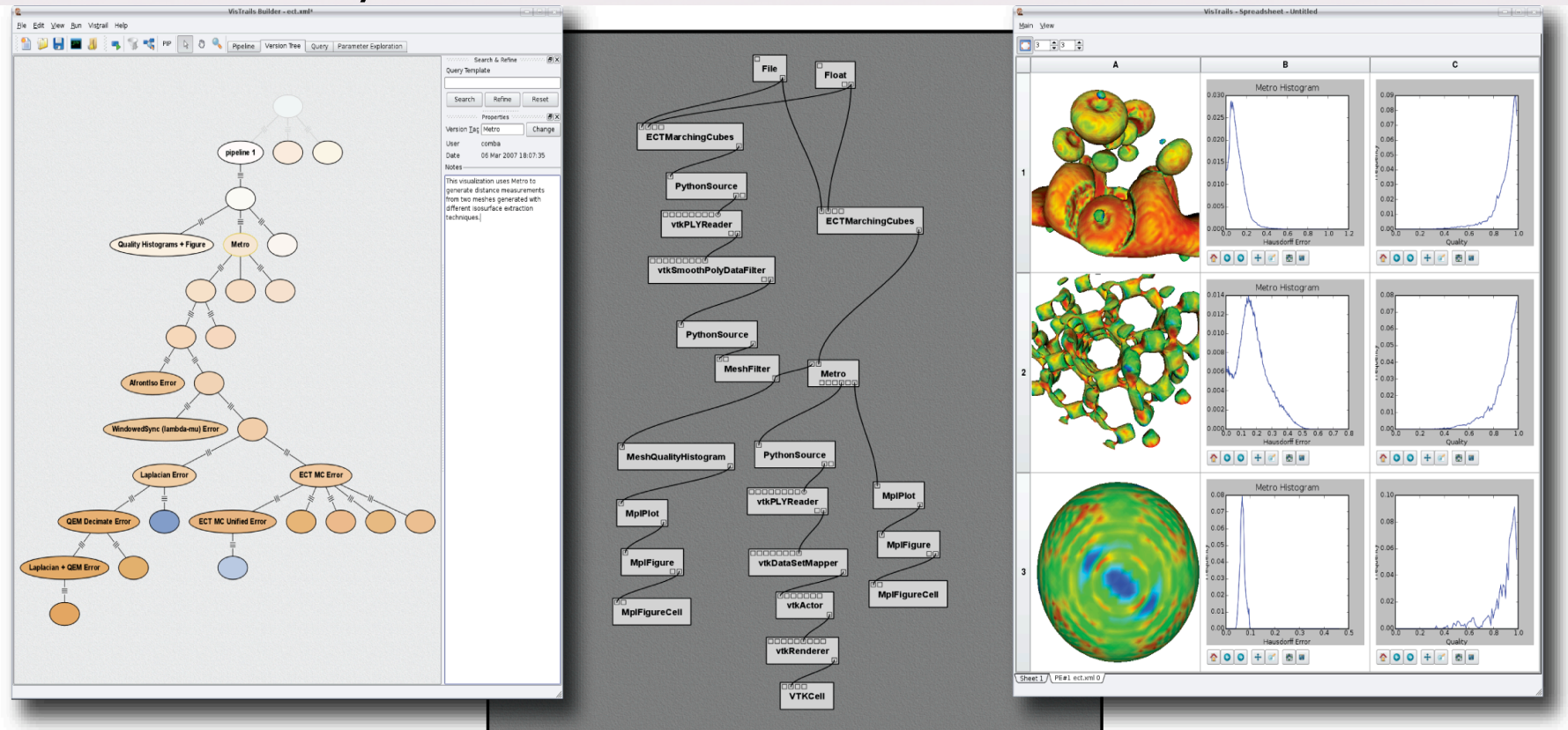
Query Result



Creating Workflows

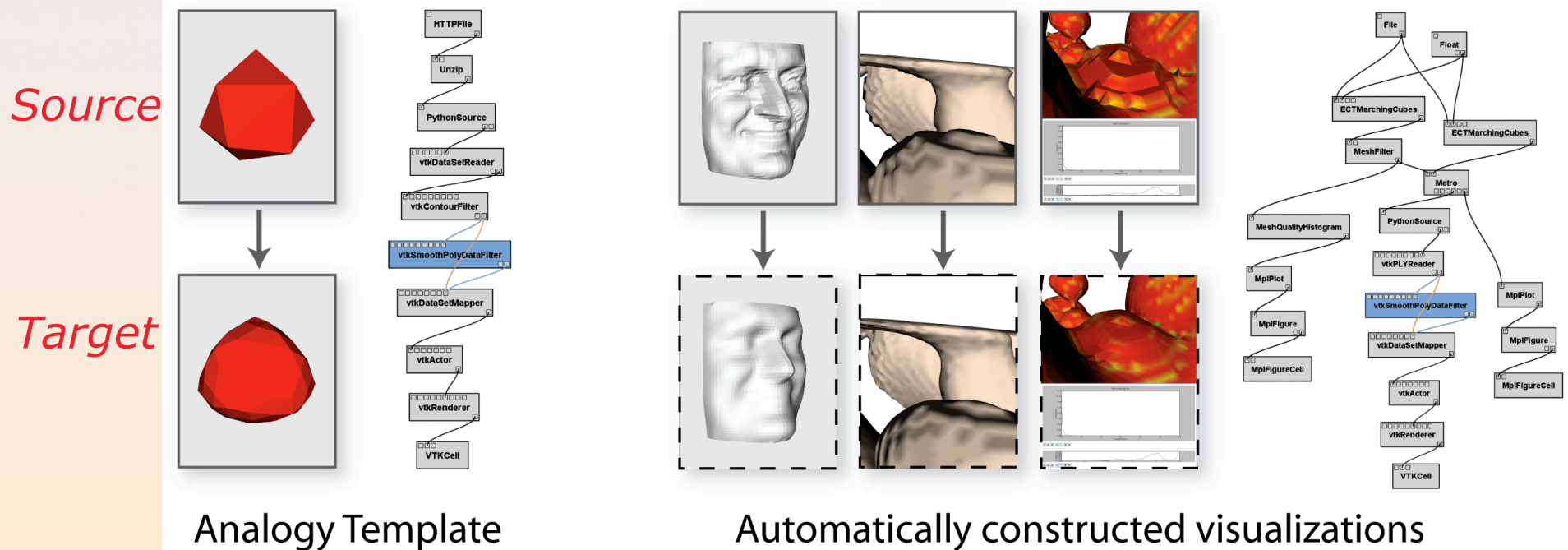
- ◆ Complex workflows are hard to create
 - Programming expertise
 - Domain knowledge
 - Familiarity with different tools

Steep learning curve

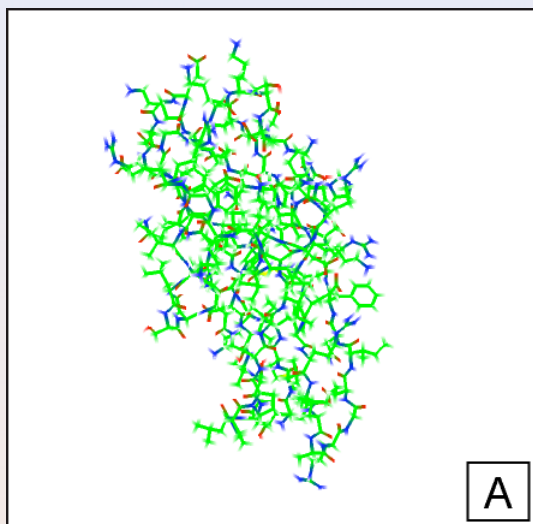


Creating Workflows by Analogy

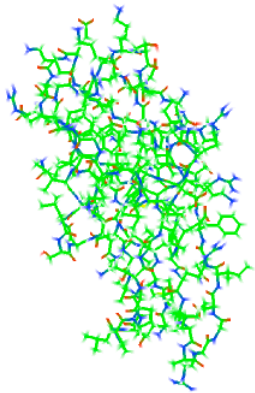
- ◆ Use the wisdom of the crowds
 - Some workflow refinements are common, e.g., change the rendering technique, publish image on the Web
- ◆ Apply refinements by analogy, automatically
[Scheidegger et al, IEEE TVCG 2007]



Creating Workflows by Analogy

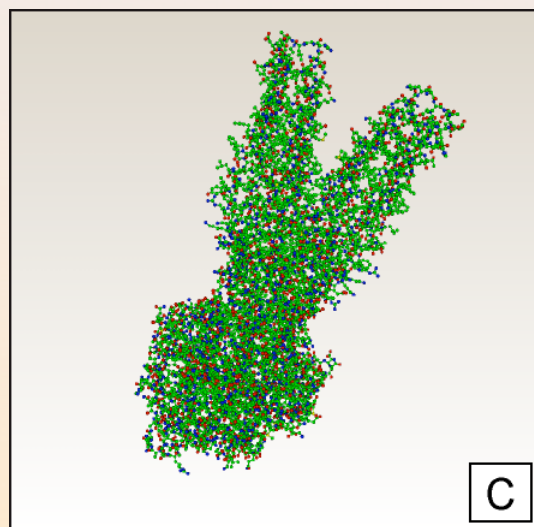


is to

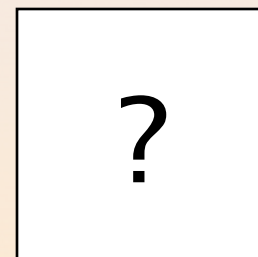
| PDB Report | |
|---|---|
|  | Protein Title NEURAL CELL ADHESION MOLECULE, MODULE 2, NMR, 20 STRUCTURES |
| | Authors P.H.JENSEN, V.SOROKA, N.K.THOMSEN, V.BEREZIN, E. BOCK, F.M.POULSEN |
| | Atom Count C: 9560 H: 15440 N: 2580 O: 2680 S: 60 |
| | Links PDB Entry |

B

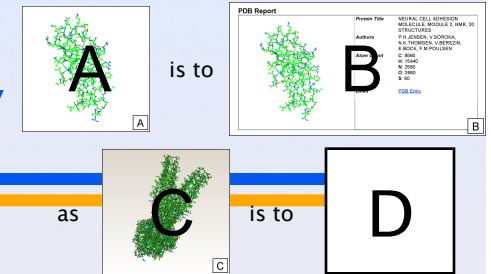
as



is to

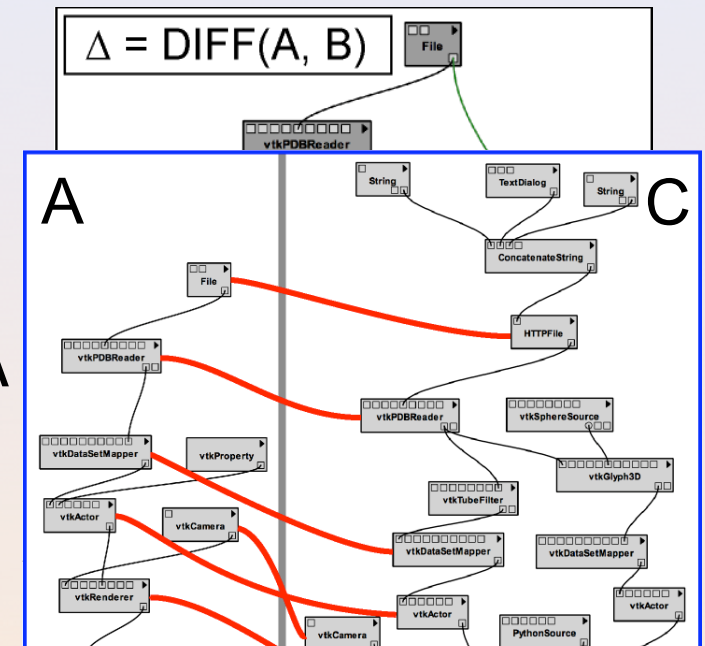


Creating Workflows by Analogy



1. Compute difference: $\Delta(A, B)$
 - Just like a patch!
 - But...

$D = \Delta(A, B) \circ C$ may not be a valid workflow
2. Find correspondences between A and C: $\text{map}(A, C)$
 - Diffuse similarity scores across the product graph $A \times C$ using Eigenvalue decompositions
3. Compute mapped difference $\Delta_{AC}(A, B) = \text{map}(A, C) \Delta(A, B)$
4. $D = \Delta_{AC}(A, B) \circ C$

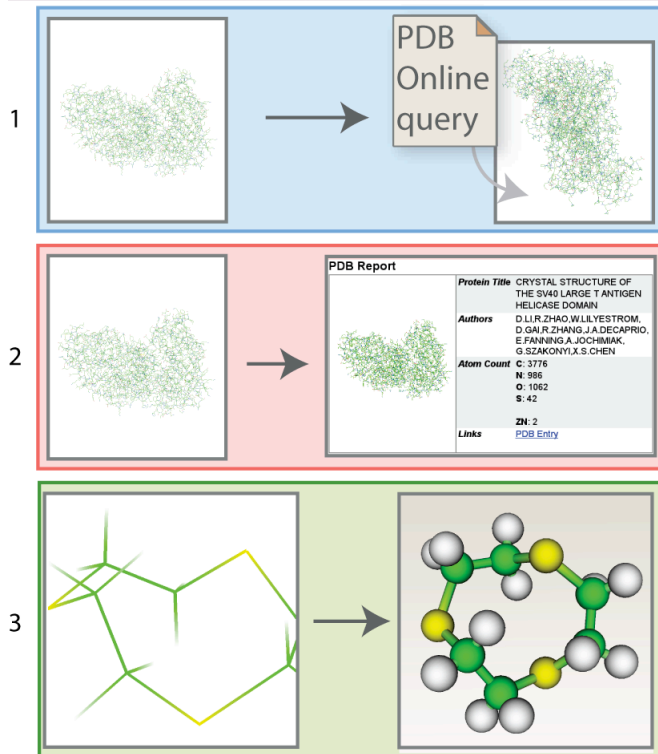


| PDB Report | |
|------------|--|
| | Protein Title STRUCTURE OF THE MULTIDRUG ABC TRANSPORTER SAV1866 FROM S. AUREUS IN COMPLEX WITH AMP-PNP |
| | Authors R.J.P.DAWSON, K.P.LOCHER |
| | Atom Count C: 5934 N: 1548 O: 1668 S: 18 |
| | C: 20 N: 12 NA: 4 O: 44 P: 6 |
| | Links PDB Entry |

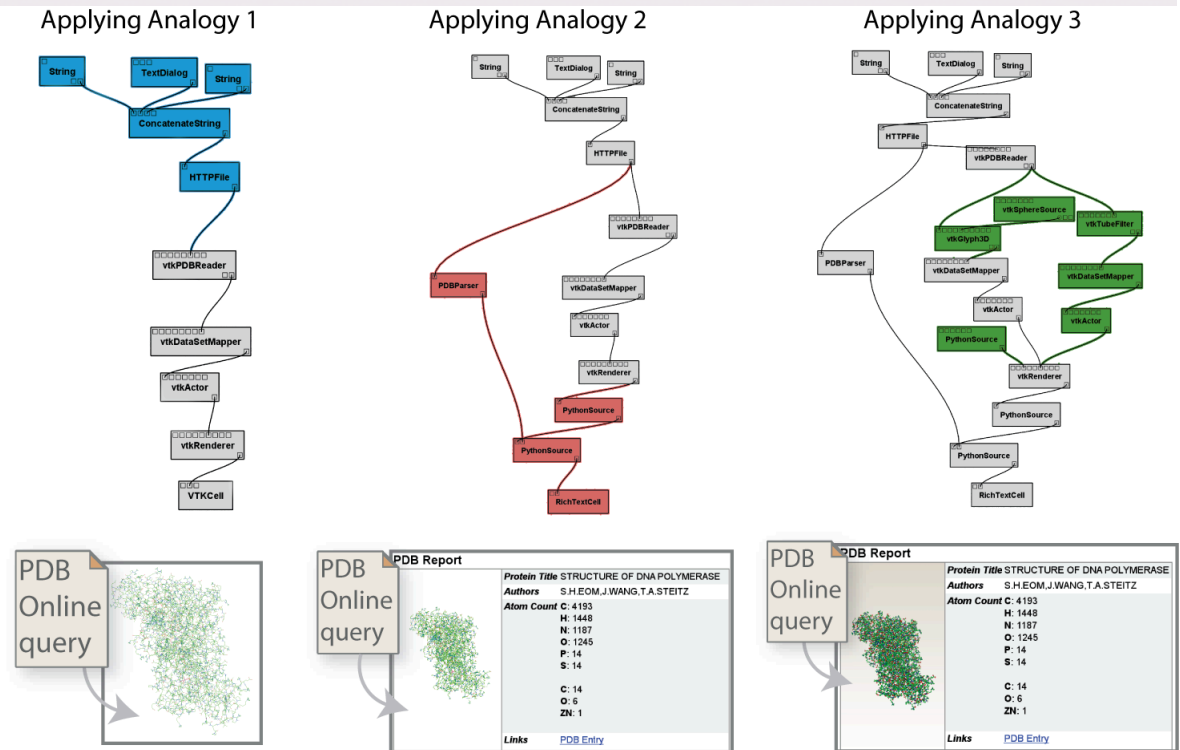
QBE and Analogies

◆ See paper:

- Querying and Re-Using Workflows with VisTrails
Carlos E. Scheidegger, David Koop, Huy Vo, Juliana Freire,
and Claudio T. Silva (**Best Paper Award at VIS 2007**)



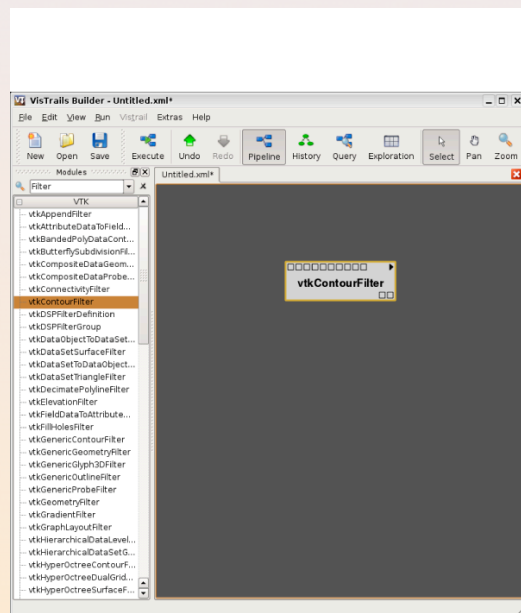
Analogy Templates



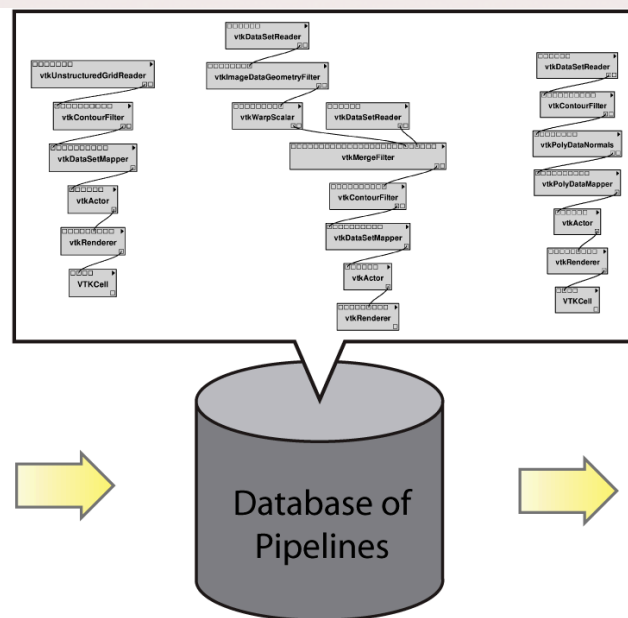
Automatically generated workflow sequence

VisComplete: A Workflow Recommendation System

- ◆ Identify graph fragments that co-occur in a collection of workflows
- ◆ Predict sets of likely workflow additions to a given partial workflow

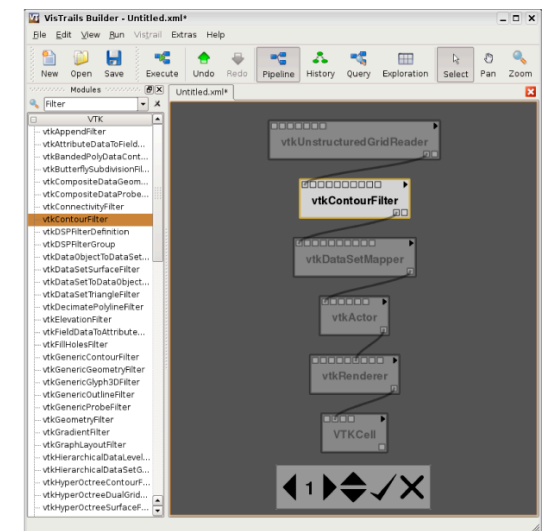


(a)



(b)

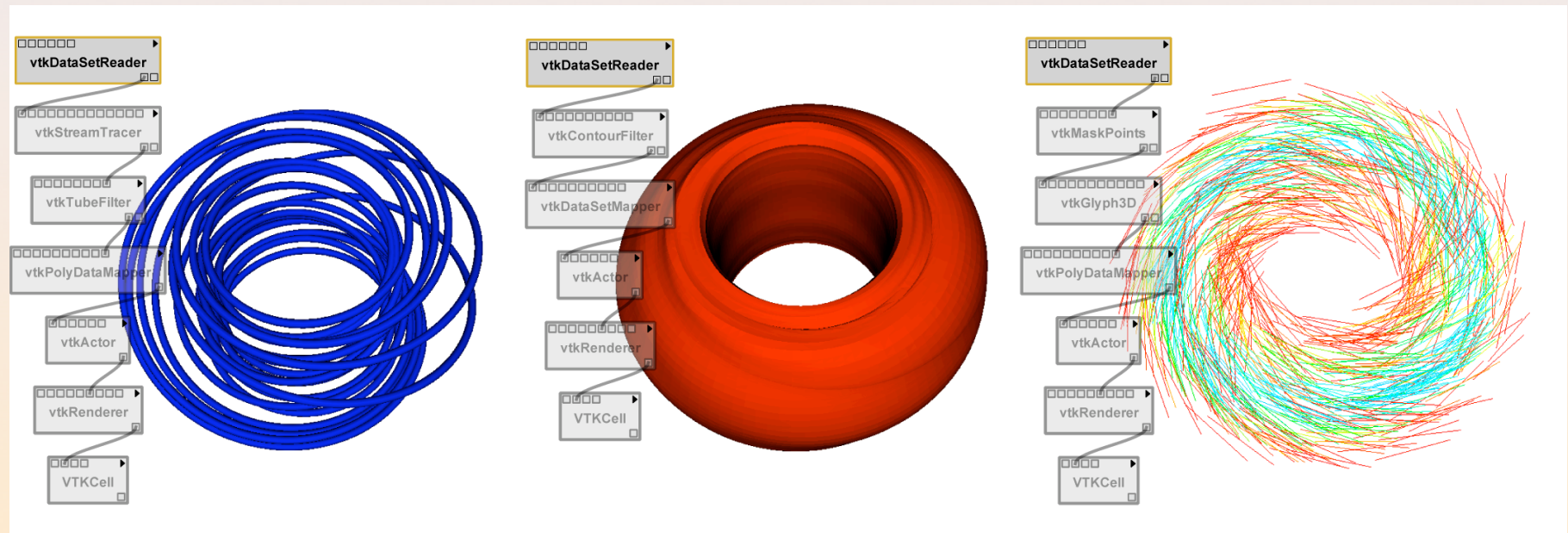
[Koop et al., IEEE Vis 2008]



(c)


VisComplete: A Workflow Recommendation System

- ◆ Similar to a Web browser suggesting URL completions
- ◆ Idea applicable to integration queries [Sarah Cohen-Boulakia et al., JBCB 2006; Talukdar et al., VLDB 2008]



VisComplete (video)

[Koop et al., IEEE Vis2008]



VisComplete:
Data-driven Suggestions for
Visualization Systems

The Provenance-Enabled Paper

- ◆ Bridge the gap between the scientific process and publications
- ◆ Results that can be reproduced and validated
 - Papers with *deep* captions
 - Encouraged by ACM SIGMOD and a number of journals
- ◆ Describe more of the discovery process: people only describe successes, can we learn from mistakes?
- ◆ Dynamic (interactive) publications
 - Evolve over time
 - Blog/wiki like=> Science 2.0
 - E.g., <http://project.liquidpub.org>
- ◆ Need tools to support this!

Provenance and Teaching (1)

- ◆ Leverage provenance to improve the way we teach CS and Science
 - www.vistrails.org/index.php/SciVisFall2007
- ◆ Lecture provenance: student can reproduce results

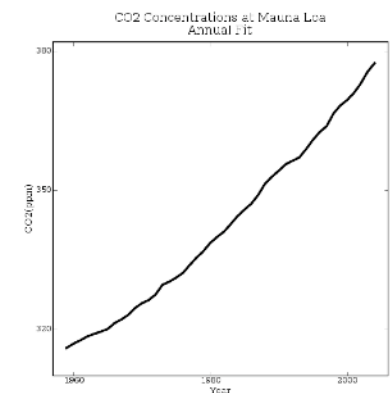
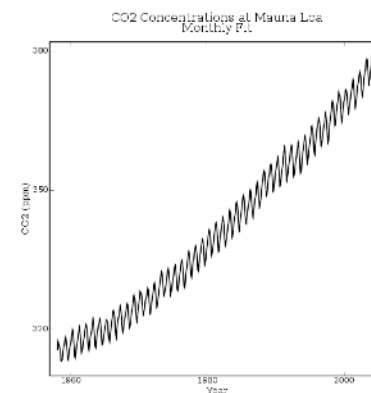
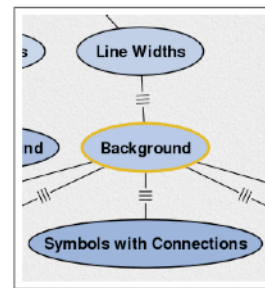
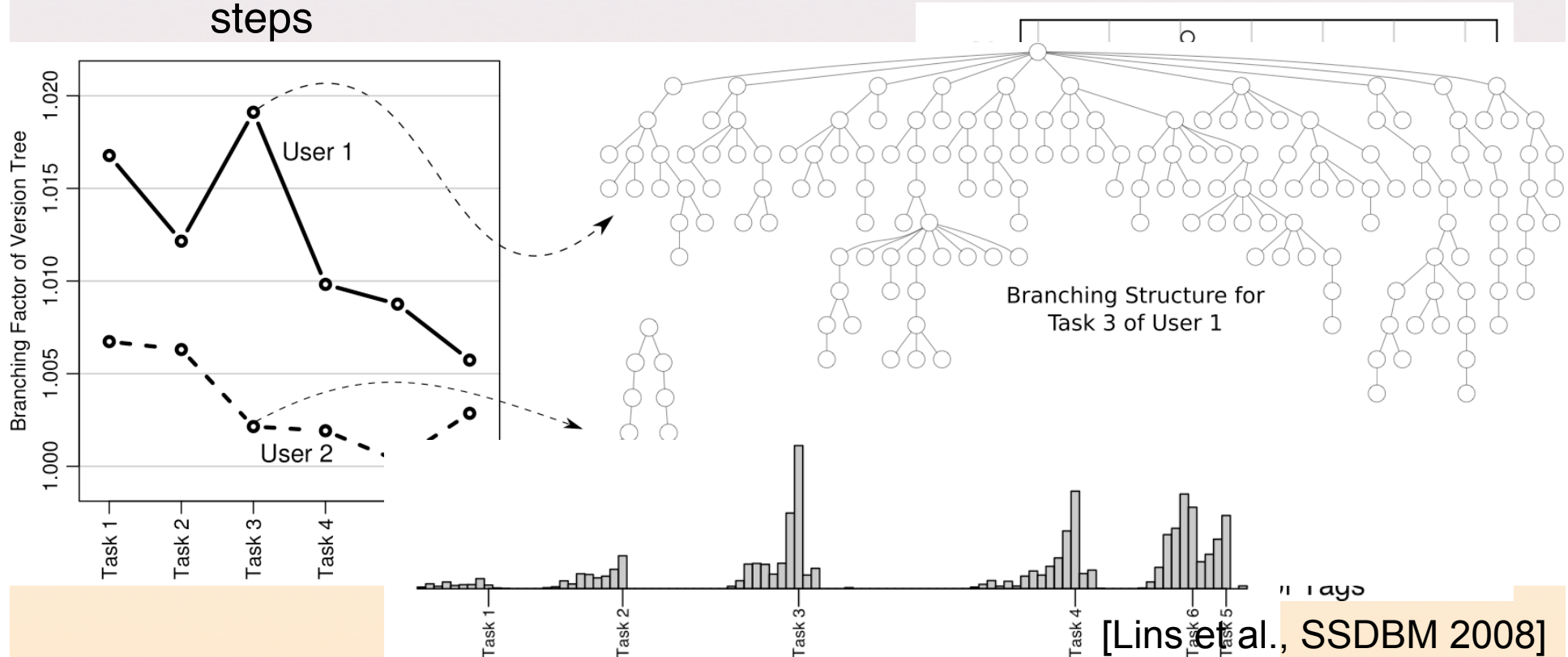


Figure 5.2: Plots of the Mauna Loa data set showing monthly measurements (left) with the yearly trend (right) using the principles for improving vision. The plot on the right is the same that was shown previously in Figure 5.1.

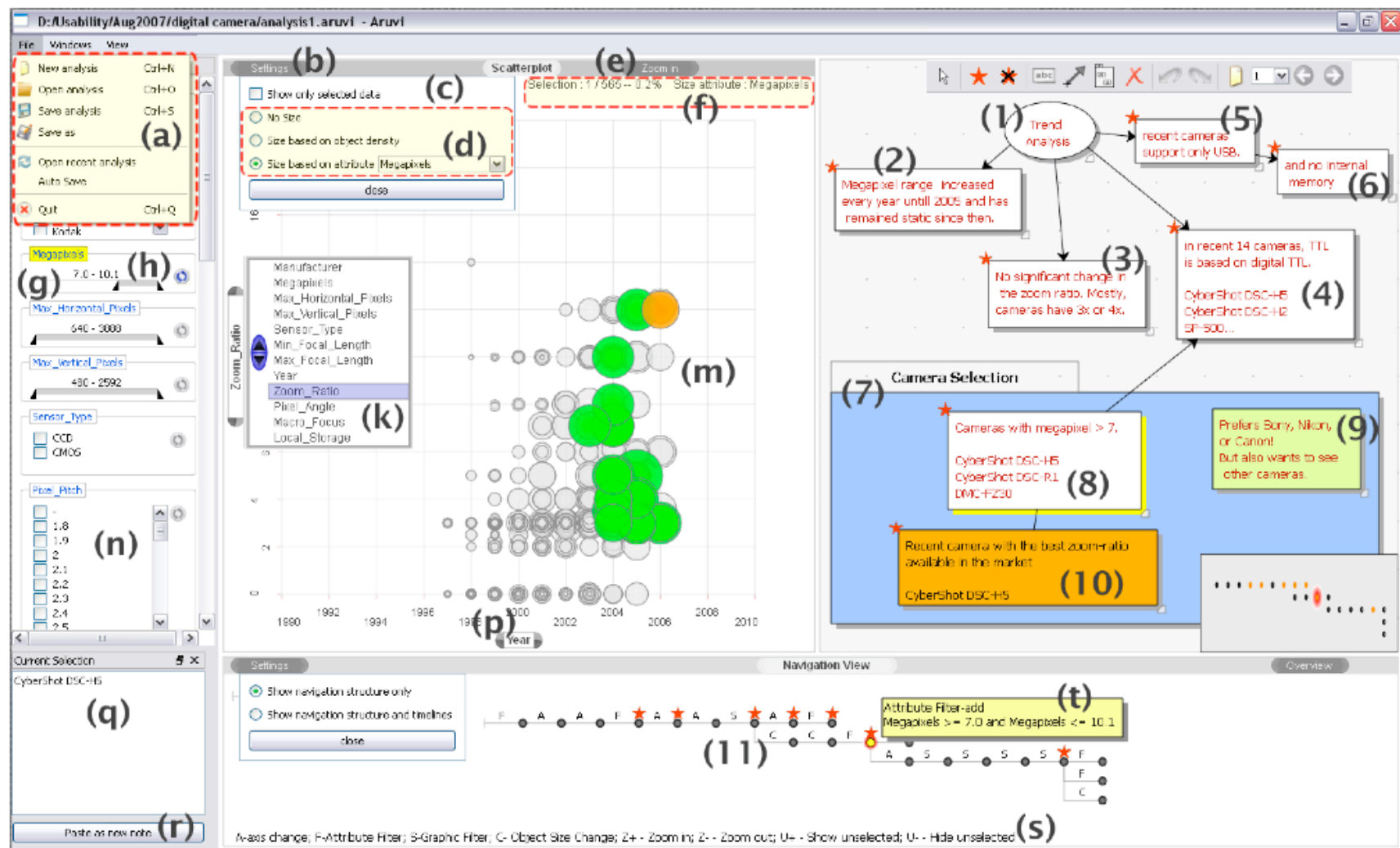
Provenance and Teaching (2)

- ◆ Workflow evolution provenance provides insights regarding
 - Task complexity and nature: number of actions; structural vs. parameter changes; task duration
 - Student confusion: large branching factor=lots of trial and error steps



Supporting the Discovery Process

Aruvi (Shrinivasan & van Wijk, 2008)

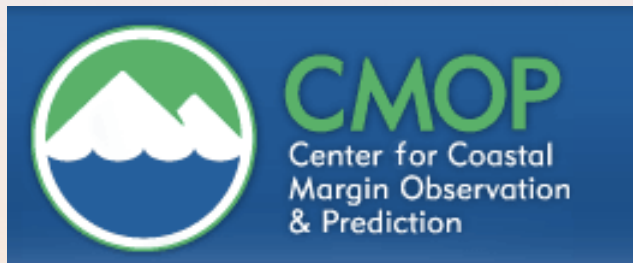


Conclusions

- ◆ We are working towards developing “infrastructure” (concepts, library, tools) that enables the development of better scientific tools!
- ◆ Adding provenance capabilities to ParaView and VisIt.
- ◆ Collaborating with Dean Williams on CDAT
- ◆ Working with Scott Klasky and the rest of SDM team on extending the SDM Dashboard with more extensive data and visualization analysis tools

Acknowledgments: Funding

- ◆ This work is partially supported by the National Science Foundation, the Department of Energy, an IBM Faculty Award, and a University of Utah Seed Grant.



More info about VisTrails

google vistrails

Or

<http://www.vistrails.org>

